Genome sequence and genetic diversity of European ash trees

Elizabeth S. A. Sollars1,2,*, Andrea L. Harper1,*, Laura J. Kelly1,*, Christine M. Sambles1,*, Ricardo H. Ramirez-Gonzalez2, David Swarbreck3, Gemy Kaithakottii4, Endymion D. Cooper5, Cristobal Uauy5, Lenka Havlickova2, Gemma Worswick1,*, David J. Studholme4, Jasmin Zohren1, Deborah L. Salmon4, Bernardo J. Clavijo5, Yi Li5, Zhesi He3, Alison Fellgett3, Lea Vig McKinney7, Lene Rostgaard Nielsen7, Gerry C. Douglas8, Erik Dahl Kjaer7, J. Allan Downie6, David Bosher9, Steve Lee10, Jo Clark11, Murray Grant4,†, Ian Bancroft3, Mario Caccamo5,12 & Richard J. A. Buggs1,13

Ash trees (genus *Fraxinus*, family Oleaceae) are widespread throughout the Northern Hemisphere, but are being devastated in Europe by the fungus *Hymenoscyphus fraxineus*, causing ash dieback, and in North America by the herbivorous beetle *Agrilus planipennis*3,4. Here we sequence the genome of a low-heterozygosity *Fraxinus excelsior* tree from Gloucestershire, UK, annotating 38,852 protein-coding genes of which 25% appear ash specific when compared with the genomes of ten other plant species. Analyses of paralogous genes suggest a whole-genome duplication shared with olive (*Olea europaea*, Oleaceae). We also re-sequence 37 *F. excelsior* trees from Europe, finding evidence for apparent long-term decline in effective population size. Using our reference sequence, we re-analyse association transcriptomic data5, yielding improved markers for reduced susceptibility to ash dieback. Surveys of these markers in British populations suggest that reduced susceptibility to ash dieback 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 a European ash (*F. excelsior*) tree generated from self-pollination of a woodland tree in Gloucestershire, UK. The sequenced tree (Earth Trust accession number 2451S) appeared free of ash dieback (ADB) when sampled in 2013 and 2014, but showed symptoms in February 2016. The haploid genome size was measured by flow cytometry as 877.24 ± 1.41 megabase pairs (Mb). Total genomic DNA was sequenced to 192× coverage (see Supplementary Table 1). We assembled the genome into 89,514 nuclear scaffolds with an N50 (the length at which scaffolds include half the bases of the assembly) of 104 kilobase pairs (kbp), 26 mitochondrial scaffolds, and one plastid chromosome (Supplementary Tables 2 and 3), where the non-N assembly constitutes 80.5% of the predicted genome size. RepeatMasker estimated 35.90% of the assembly to be repetitive elements, with long terminal repeat retrotransposons predominating (Supplementary Table 4). Compared with other eudicot genomes of similar size4,5 this repeat content is low. The 17% of the assembly composed of undetermined bases probably contains additional repeats; 27% of reads that do not map to the assembly align to ash repeats (Supplementary Table 5). We generated approximately 160 million RNA sequencing (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 (version of genome (v)2.3)5, 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, 81.80% showed high similarity to plant proteins (>50% high-scoring segment pair coverage) (Supplementary Table 8), 97.05% had matches in the non-redundant databases (excluding hits to ash), 82.74% generated hits to InterPro signatures and 78.09% were assigned Gene Ontology terms. We also identified 107 microRNA (miRNA), 792 transfer RNA (tRNA) and 51 ribosomal RNA (rRNA) genes.

Past whole-genome duplication events are commonly inferred from the distributions of pairwise synonymous site divergence (Ks) within paralogous gene groups8. We plotted these for ash and six other plant species (Fig. 1a and Supplementary Table 9). Ash and olive shared a peak near Ks = 0.25, suggesting an Oleaceae-specific whole-genome duplication. A peak near Ks = 0.6 shared by ash, olive, monkey flower and tomato but not by bladderwort, coffee and grape does not fit a common origin hypothesis, unless bladderwort has an accelerated substitution rate and the tomato peak is not restricted to the Solanales as evidenced previously4. Synteny analysis between ash and monkey flower did not provide conclusive evidence for shared whole-genome duplication (Extended Data Fig. 2). Duplicated genes in the ash genome that were not locally duplicated (that is, within ten genes of each other in our assembly) show no significantly enriched Gene Ontology terms at a false discovery rate level of 0.05. By contrast 1,005 locally duplicated genes showed significant enrichment of terms relating to oxidoreductase, catalytic and monoxygenase activity compared with all other genes, suggesting evolution of secondary metabolism by local duplications.

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

To investigate genomic diversity in F. excelsior, we sequenced 37 ash trees from central, northern and western Europe (Fig. 2 and Supplementary Table 11), to an average of 8.4× genome coverage by trimmed and filtered reads. Together with reads from Danish ‘Tree35’ (http://oadb.tsl.ac.uk/), these were mapped to the reference genome. We found 12.48 million polymorphic sites with a variant of high confidence in at least one individual (quality > 300 using freebayes10): we refer to these as the ‘genome-wide SNP set’ in the ‘European Diversity Panel’. Of these, 6.85 million (54.88%) occur inside or within 5 kbp of genes (Supplementary Table 12). We found 259,946 amino-acid substitutions and 71,513 variants that affect stop or start codons, or splice sites.

We selected 23 amino-acid variants, and 26 non-coding variants from the ‘genome-wide SNP set’ with a range of call qualities for validation using KASP: individual genotype calls with quality greater than 300 have a false-positive rate of 6% and those with quality greater than 1,000 have a false-positive rate of zero (Supplementary Table 13). We ran a more stringent variant calling restricted to regions of the genome with between 5× and 30× coverage in all 38 samples. These totalled 20.6 Mbp (2.3% of the genome), within which 529,812 variants were called with CLC Genomics Workbench. Of these, 394,885 were bi-allelic single nucleotide polymorphisms (SNPs) with minimum allele frequency above 0.05, which we refer to as the ‘reduced SNP set’. We also found about 31,300 singleton simple sequence repeat (SSR) loci in the ash genome, and designed primers for 664 (Supplementary Data 1). In a sample of 366 of these, 48% were polymorphic in the European Diversity Panel sequences. We PCR tested 48 of these in multiplexes with European Diversity Panel genomic DNA and found that 41 amplified successfully (Supplementary Data 1).

We analysed population structure of the European Diversity Panel using a plastid haplotype network; STRUCTURE11 runs on genomic SNPs and principal component analysis (PCA) of the ‘reduced SNP set’ (Fig. 2a–d and Extended Data Fig. 3). Clearest differentiation was found in the plastid network, with four distinct haplotype groups each separated from each other by at least 20 substitutions. One group was more frequent in Great Britain than on the continental Europe. The second and third principal components of the PCA corresponded to the plastid data somewhat (Fig. 2c). Previous analyses of SSRs in plastids identified variants unique to the British Isles and Iberia12. Linkage disequilibrium in the European Diversity Panel decayed logarithmically, with an average $r^2$ of 0.15 at 100 bp between SNPs, reaching an $r^2$ of 0.05 at ~40 kbp (Fig. 2e). This is similar to long-range linkage disequilibrium estimates found in Populus tremuloides13. An apparent long-term effective population size decline of F. excelsior in Europe was shown by analyses based on heterozygosity in the reference genome (using pairwise sequentially Markovian coalescent (PSMC)14, Fig. 2f). Such patterns may also reflect a complex history of population subdivision in ash15.

We used associative transcriptomics to predict ADB damage in Great Britain. We used the full coding DNA sequence (CDS) models from our genome annotation as a mapping reference for previously generated16 RNA-seq reads from 182 Danish ash accessions (‘Danish Scored Panel’) that have been exposed to H. fraxineus, and scored for damage (Supplementary Data 2). This yielded 40,133 gene expression markers (GEMs; Supplementary Data 3) and 394,006 SNPs (Supplementary Data 4). Twenty GEMs were associated with ADB damage scores, including eight MADS-box proteins, and two cinnamoyl-CoA reductase 2 genes that may be involved in the hypersensitive response (Supplementary Data 5). Four assays representing the top five GEMs were applied to 58 Danish accessions (‘Danish Test Panel’) to validate the top markers. Results were combined into a single predicted damage

---

**Figure 1** Gene sharing within and among plant genomes.

a. Distribution of $K_s$ values between paralogous gene pairs within the genome of ash (F. excelsior), tomato (Solanum lycopersicum), coffee (Coffea canephora), bladderwort (Utricularia gibba), grape (Vitis vinifera) and monkey flower (Mimulus guttatus), and transcriptome of olive (O. europaea). b. Venn diagram of gene sharing by five asterid species. c. Venn diagram of gene sharing by six woody species. Numbers in parentheses are the total number of OrthoMCL groups found for that species; numbers in intersections show the total number of groups shared between given combinations of taxa.
We also examined expression of the top five GEM loci using reads per kilobase pair per million aligned reads (RPKM) values from our shotgun Illumina read data for the reference tree (Extended Data Fig. 4), comparing these with RPKM values from the Danish Scoring Panel. Expression patterns in the reference tree were highly correlated with those of the most susceptible Danish quartile ($r^2 = 0.995$, $P < 0.001$), but not with the least susceptible ($P = 0.24$), consistent with observations that the reference tree is now succumbing to the disease. We correlated the expression of all 20 top GEM markers in leaf, flower, cambium and root transcriptomes of the parent of the reference tree. This revealed that leaf expression levels were positively correlated with those in the cambium ($r^2 = 0.65$, $P < 0.001$) and flower ($r^2 = 0.38$, $P = 0.0041$), but not with the root ($P = 0.3594$).

We identified putative orthologues of the five GEM loci using our OrthoMCL results (Supplementary Data 5) and BLAST searches of GenBank, and conducted maximum likelihood and Bayesian analyses of relevant hits (Extended Data Fig. 5).

**Figure 3 | Predicted ADB damage scores in Great Britain and Denmark.** Map points are scaled by hue (high predicted damage scores in brown, low in green) and plotted according to the geographical origin of the parent trees of the British Screening Panel ($n = 130$) and the Danish Test Panel ($n = 58$). Single leaf samples taken from grafts of each individual tree were used for predicting damage scores. Inset: damage predictions for the Danish Test Panel ($n = 58$) correlated with log mean observed damage scores from 2013 to 2014 ($r^2 = 0.25$, $P = 6.9 \times 10^{-8}$).

We identified putative orthologues of the five GEM loci using our OrthoMCL results (Supplementary Data 5) and BLAST searches of GenBank, and conducted maximum likelihood and Bayesian analyses of relevant hits (Extended Data Fig. 5). FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 clustered into the SVP/StMADS11 group16 of type II MADS-box genes. FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1 clustered into the SOC1/TM3 group of type II MADS-box proteins16,17. Both groups have roles in flower development18–21, and appear to be involved in stress response in *Brassica rapa*22. Many genes involved in regulation of flowering time in *Arabidopsis thaliana* are involved in controlling phenology in perennial trees species23, and genes belonging to the SVP/StMADS11 clade have potential roles in growth cessation, bud set and dormancy25. In *A. thaliana*, AGL22/SVP may be required for age-related resistance24.

One mechanism by which transcriptional cascades, such as those involving MADS box genes, might be involved in tolerance or resistance to pathogens is via modulation of secondary metabolite concentrations. For five high-susceptibility and five low-susceptibility Danish trees, we profiled methanol-extracted leaf samples by liquid chromatography/mass spectrometry on a quadrupole time-of-flight mass spectrometer. Partial least squares discriminant analysis (PLS-DA) clearly discriminated high- and low-susceptibility trees (Fig. 4a). By using accurate mass to identify the chemical nature of discriminant features, we found greater abundance (Fig. 4b) of iridoid glycosides (for details see Extended Data Figs 6–9 and Supplementary Data 9) in genotypes with high susceptibility to ADB than in low-susceptibility genotypes.
A tandem mass spectrometry (MS/MS) fragmentation network identified several product ions expected from fragmentation of iridoid glycosides (Fig. 4c). Iridoid glycosides are a well-known anti-herbivore defence mechanism in the Oleaceae. They can also enhance fungal growth defence mechanism in the Oleaceae. They can also enhance fungal resistance. Iridoid glycosides are a well-known anti-herbivore defence mechanism in the Oleaceae.

**Figure 4** | Putative iridoid glycosides as discriminatory features between *F. excelsior* genotypes with differential susceptibility to ADB. 

a. Multivariate analysis PLS-DA score plot of metabolic profiles of five high-susceptibility and five low-susceptibility trees (n = 3 per genotype). 

b. Box-plots from these profiles showing normalized (internal standard) intensity (log2 transformed) of five discriminatory features observed in negative mode; m/z and retention time (RT) are given for each feature. 

c. Fragmentation network of discriminatory features, highlighted in black (positive mode) and grey (negative mode). Each product ion is labelled with its size (m/z), also depicted by its circle size. Blue shading increases with the number of times each ion is present in the precursor discriminatory features. Product ions not shared among precursors are shown as unlabelled tips. The edges are in shades of red on the basis of retention time; the paler the colour the earlier the retention time. Those fragment masses shaded in green have been previously reported from fragmentation of iridoid glycosides.
Letter

Supplementary Information is available in the online version of the paper.

Acknowledgements Eurofins MWG provided a discounted service for Illumina and 454 sequencing of the reference genome, funded by Natural Environment Research Council (NERC) Urgency Grant NE/K01112X/1 to R.J.A.B. The associative transcriptomic and metabolomic work was part of the ‘Nornex’ project led by J.A.D. funded jointly by the UK Biotechnology and Biological Sciences Research Council (BBSRC) (BBS/E/J/000CA5323) and the Department for Environment, Food & Rural Affairs. The Earleham Institute, Norwich, UK, sequenced ‘Tree 35’ funded by ‘Nornex’ and the European Diversity Panel funded by the Earleham Institute National Capability in Genomics (BB/J010375/1) grant. W. Crowther assisted with DNA extractions for the KASP assay; The John Innes Centre contributed KASP analyses. J. F. Miranda assisted with RNA extractions and quantitative PCR with reverse transcription (qRT-PCR) at the University of York. H. V. Florance, N. Smirnoff and the Exeter Metabolomics Facility developed metabolomic methods and ran samples, and T. P. Howard helped with statistics. L. J. K. and R. J. A. B. were partly funded by Living with Environmental Change (LWEC) Tree Health and Plant Biosecurity Initiative - Phase 2 grant BB/L012162/1 to R. J. A. B., S. L. and P. Jeppson funded jointly by a grant from the BBSRC, Defra, Economic and Social Research Council, and the Forestry Commission. R.J.A.B. also received support from the Tree Health and Plant Biosecurity Initiative. G.W. was funded by Teagasc Walsh Fellowship 2014001 to R.J.A.B. and G.C.D. E.C. was funded by a Marie Skłodowska-Curie Individual Fellowship ‘FraxiFam’ (grant agreement 660003) to E.D.C. and R.J.A.B. E.S.A.S. and J.J.C. were funded by the Marie Skłodowska-Curie Initial Training Network INTERCROSSING. J.A.D. received a John Innes Foundation fellowship. We thank A. Jöecker for supervising E.S.A.S. at Giangen and for helpful discussions. R.H.R.G. is supported by a Norwich Research Park PhD Studentship and Earleham Institute Funding and Maintenance Grant. This research used Queen Mary’s MidPlus computational facilities, supported by QMUL Research-IT and funded by Engineering and Physical Sciences Research Council grant EP/K000128/1 and NERC EOS Cloud. D.J.S. acknowledges the support of BBSRC grant BB/N021452/1, which partly supported M.G., C.M.S. and D.J.S. during this work.

Author Contributions 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 the project and directed work on the reference genome. E.S.A.S. assembled the reference genome and organellar genomes, and analysed gene and genome duplications, European population structure and past effective population sizes. L.J.K. extracted high molecular mass DNA for the European Diversity Panel and conducted repetitive element, OrthoMCL and phylogenetic analyses. G.W. conducted SSR analyses. J.J.C. extracted high molecular mass DNA and RNA for the reference genome. E.D.C. analysed genome duplication in the reference genome. D.S. and G.K. performed bioinformatic analyses to annotate the reference genome. M.C. conceived and conducted the analyses; C.M.S. developed methodology; and D.L.S. processed and extracted samples and ran the mass spectrometer.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.A.B. (rbuggs@qmul.ac.uk).

Reviewer Information Nature thanks P. Dorrestein, S. Jansson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/
METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Tree material. Reference tree. In 2013 twig material was collected from tree 2451S growing at Paradise Wood, Earth Trust, Oxfordshire, UK. This tree was produced from self-pollination of a hermaphroditic F. excelsior tree growing in woodland in Gloucestershire (latitude 52.020592, longitude -1.832804), UK, in 2002 as part of the BRC F. excelsior project.24 The parent tree was one of 19 trees that produced seed from self-pollination, and had lower heterozygosity at four microsatellite loci than the other 18 trees (D.B., unpublished observations). DNA was extracted from bud, cambial and wood tissues using CTAB23 and Qiagen DNeasy protocols. RNA was extracted using the Qiagen RNAeasy protocol from leaf tissue of tree 2451S and from leaf, cambium, root, and flower tissue of its parent tree in Gloucestershire.

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

British Screening Panel. In 2015, freshly flushed leaf material was collected from a clonal seed orchard of F. excelsior growing at Paradise Wood, Earth Trust, Oxfordshire, UK, for RNA extraction and complementary DNA (cDNA) synthesis as in ref. 3. Single whole leaves were harvested from four ramets of each of 130 ash trees selected from phenotypically superior parents throughout Britain, which had been cloned by grafting.

2451S DNA sequencing and genome assembly. The genome size of 2451S was estimated by flow cytometry with propidium iodide staining of nuclei, using leaf tissue co-chopped with an internal standard using a razor blade. Three preparations were made, two with *Petroselinum crispum* 'Curled Moss' parsley as standard (2C genome size = 4.50 pg)15 and one with *S. foetidissimum* Stupicke polni rane (2C = 1.96 pg)15 as standard. The Partec CyStain Absolute Prot protocol was used (Partec, Germany). Each preparation was measured six times, with the relative fluorescence of over 5,000 particles per replicate recorded on a Partec Cyflow SL3 (Partec, Germany) flow cytometer fitted with a 100-mW green solid state laser (Partec, Germany). The resulting histograms were analysed with the Flow-Max software (version 2.4, Partec). The measurement with the tomato internal standard was used as the best estimate of genome size, because the tomato genome size is closest to that of 2451S, yielding a more accurate result.

Genomic DNA of 2451S was sequenced using the following methods: (1) HiSeq 2000 (Illumina, San Diego, California, USA) at Eurofins, Ebersberg, Germany, with 100 bp reads and shotgun libraries with fragment sizes of 200 bp, 300 bp, and 500 bp, and 10,000 jumping distance libraries with 3 kbp, 8 kbp, 20 kbp, and 40 kbp insert sizes, generating 188× genome coverage; (2) 454 FLX+ (Roche, Switzerland) at Eurofins with shotgun libraries and maximum read length of 1.763 bp and mean length of 642 bp giving 4.3× genome coverage; and (3) MiSeq (Illumina, San Diego, California) at the Earlham Institute, Norwich, UK, with 300 bp paired-end reads from a Nextera library with ~3 kbp insert size, giving 16× genome coverage (see Supplementary Table 1). We assembled and released five genome assembly versions over the course of 3 years, details of which can be found in Supplementary Table 3. The most recent version assembled first into 235,463 contigs with a total size of 663 Mbp and an *N₅₀* of 5.7 kbp (Supplementary Table 2), and after scaffolding and removing organellar scaffolds, the assembly comprised 89,487 scaffolds totalling 867 Mbp (17% ‘N’) with an *N₅₀* of 104 kbp (Supplementary Table 2). The plastid genome was assembled separately into one circular contig of 155,498 bp, including an inverted repeat region of approximately 25,700 bp. The mitochondrial genome initially assembled into 296 contigs totalling 232 kbp. After several rounds of contig extension using overlaps of mapped 454 reads, the final assembly consisted of 26 contigs with an *N₅₀* of 64.3 kbp.

All illumina reads from 2451S were trimmed using CLC Genomics Workbench (QIAGEN Aarhus, Denmark) versions 6–8 (depending on when the data were received) to a minimum quality score of 0.01 (equivalent to Phred quality score of 20), a minimum length of 50 bp, and were trimmed of any adaptor and repetitive telomere sequences. The MiSeq Nextera reads were also run through FLASH26 to merge overlapping paired reads, and NextClip46 to remove adaptor sequences, both used with default parameters. Roche 454 reads were trimmed to a minimum Phred score of 0.05, and minimum length of 50 bp. De novo assembly was performed with the CLC Genomics Workbench, using the 200 bp, 300 bp, 500 bp and 5 kbp insert size Illumina library reads to build the De Bruijn graphs. The remaining Illumina reads and the 454 reads were used as ‘guidance only reads’ to help select the most supported path through the De Bruijn graphs. A word size (k-mer, a substring of length k in DNA sequence data) of 50 and maximum bubble size of 5,000 were used to assemble the reads into contigs with a minimum length of 500 bp. Contigs were then scaffolded with the stand-alone tool SSPACE27 Basic version 2.0 using all paired illumina reads, with the ‘k’ parameter (number of mapped paired reads required to join contigs) set to 7. Gaps in the scaffolds were closed using the GapCloser version 1.12 program using all paired reads (except for long jumping distance libraries), with pair_num_cutoff parameter set at 7. Four hundred and fifty-four reads were mapped to the assembly and used to join overlapping scaffolds using the jellyfish.py script from PBSuite48 version 14.7.14 with the following blast parameters: -minMatch 11 -minPctIdentity 70 -bestn 1 -nCandidates 10 -maxScore -500 -noSplitSubreads. Contigs57 were removed from the assembly because it was aligned fully to the Phix bacterial genome, indicating it derived from the Phix control library added to Illumina sequencing runs.

To assemble the plastid and mitochondrial genomes, high read depth 50bp k-mers were extracted from the 200, 300 and 500 bp read libraries. Jellyfish39 version 2.1.1 was used to count the depth for each k-mer, and these values were plotted in a scatterplot to identify peaks that could correspond to the organellar genomes. Every k-mer over 600× coverage was used in a BLAST search against the NCBI non-redundant (nr) database with a filter allowing only plant sequences; k-mers were then extracted on the basis of whether their first hit contained a ‘mitochondrion’ or ‘plastid/chloroplast’ related description. Reads from the 200, 300 and 500 bp libraries were then filtered against the k-mer sets, and were kept if the first and last 50 bp matched k-mers from the extracted sets (reads at most 90 bp long). Each set of reads (mitochondrial and plastid) were then assembled de novo using the CLC Genomics Workbench. The plastid genome assembled initially into two contigs, which were joined using an alignment to the *O. europaea* plastid genome (GenBank accession number NC_015401.1), with the inverted repeat region being identified also. Reads from the 454 library were mapped to the assembly to check the sequence and especially the join region. The mitochondrial genome assembled first into 296 contigs. To fill in gaps and join the contigs together, 454 reads were mapped against the assembly and contig ends were extended using the Extend Contigs tool in the CLC Genome Fiming Module. The Join Contigs tool was then used to join overlapping ends together, and 454 reads were mapped to the resulting assembly to check any joined regions. Using this method of ‘Map-Extend-Join’ iteratively (approximately ten times in total), a more contiguous assembly of 26 contigs was obtained.

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

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

De novo identification of repeat families from unassembled reads. Individual 454 reads and Illumina read pairs from the 500 bp insert library (after adaptor trimming, but before any further quality control or filtering; see above) were used for de novo repeat identification. Reads were quality filtered and trimmed using the FASTX-Toolkit version 0.13.10 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Using fastx_trimmer, the first 10 bp of all reads (454 and Illumina) were removed (owing to skewed base composition). The 454 reads were clipped to a maximum of 250 bp and Illumina reads to a maximum of 90 bp; all shorter reads were removed using a custom Perl script. Reads were then quality filtered with the fastq_quality_filter tool to retain only those where 90% of bases had a Phred score of at least 20. Exact duplicates (which are probably artefacts from the emulsion PCR26) were removed from the 454 reads using the fastx_collapser tool.

A complete set of quality filtered and trimmed 454 reads (3,330,483) was used as input for the RepeatExplorer pipeline on Galaxy47, with a minimum of 138 bp overlap for clustering and a minimum of 100 bp overlap for assembly. All clusters containing at least 0.01% of the input reads were examined manually to identify clusters that required merging (that is, where there was evidence that a single repeat family had been split over multiple clusters). Clusters were merged if they met the following three criteria: (1) they shared a significant number of similarity hits (for example, in a pair of clusters, 10% of the reads in the smaller cluster had BLAST hits to reads in the larger cluster); (2) they were the same repeat type (for example, LINEs), (3) they could be merged in a logical position (for example, one cluster containing conserved domains these domains would be joined in the correct order). The re-clustering pipeline was run with a minimum of 100 bp overlap for assembly; merged clusters were examined manually to verify that all domains were in the correct orientation.

Quality filtered and trimmed Illumina reads were paired using the FASTA interlaver tool (version 1.0.0) in RepeatExplorer, resulting in 111,230,011 pairs;
unpaired sequences were discarded. An initial run of RepeatExplorer with a sample of 100,000 read pairs was performed to obtain an estimate of the maximum number of reads that could be handled by the pipeline. A random sample of 3.5 million read pairs was then taken using the sequence sampling tool (version 1.0.0) in RepeatExplorer and used as input for the clustering pipeline, which further randomly subsampled the reads down to 370,186 pairs. The pipeline was run with a minimum of 50 bp overlap for clustering and a minimum of 36 bp overlap for assembly. Clusters containing at least 0.01% of the input reads were merged if \( k_{\text{c}} \) passed the 0.2 cut-off (for clusters \( x \) and \( y \), \( k_{\text{c}} \) is defined as \( k_{\text{c}} = 2W(n_1 + n_2) \) where \( W \) is the number of read pairs shared between clusters \( x \) and \( y \), and \( n_1 \) is the number of reads in cluster \( x \) which does not include the other read from its pair within the same cluster); clusters that passed this threshold but which had no similarity hits to each other were not merged. The re-clustering pipeline was run with a minimum of 36 bp overlap for assembly.

Repeat families identified by RepeatExplorer were annotated according to the results of BLAST searches to the Viridiplantae RepeatMasker library, with a value cut-off of 1 (ref. 44). Any repeat families matching the custom library of ash repeats using BLASTN as described were removed before downstream analyses (see below).

Identification of repetitive elements from the genome assembly. *De novo* identification of repetitive elements from the assembled ash genome sequence was conducted with RepeatModeler version 1.0.7 (http://www.repeatmasker.org/RepeatModeler.html) using RMBlast as the search engine. All unannotated (‘unknown’) repeat families from the RepeatModeler library were searched against a custom BLAST database of organellar genomes (see above) using BLASTN with an E value cutoff of 1 \( \times 10^{-10} \) and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed.

To prevent any captured gene fragments within repetitive element families causing the masking of protein coding genes within the ash assembly, the custom repeat libraries were pre-masked using the TAIR10 CDS data set (TAIR10_cds_20101214_updated; downloaded from http://www.arabidopsis.org). First, RepeatModeler was run with a minimum of 50 bp overlap for clustering and a minimum of 36 bp overlap for assembly. Clusters containing at least 0.01% of the input reads were merged if \( k_{\text{c}} \) passed the 0.2 cut-off (for clusters \( x \) and \( y \), \( k_{\text{c}} \) is defined as \( k_{\text{c}} = 2W(n_1 + n_2) \) where \( W \) is the number of read pairs shared between clusters \( x \) and \( y \), and \( n_1 \) is the number of reads in cluster \( x \) which does not include the other read from its pair within the same cluster); clusters that passed this threshold but which had no similarity hits to each other were not merged. The re-clustering pipeline was run with a minimum of 36 bp overlap for assembly.

Repeat families identified by RepeatExplorer were annotated according to the results of BLAST searches to the Viridiplantae RepeatMasker library, to a value cut-off of 1.

**Gene annotation**.

Expression of transposable elements in F. excelsior. To test whether the about 18% of 'Gene annotation').

The filtered TAIR10 CDS data set was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and Illumina) and the library from RepeatMasker using RepeatMasker version 4.0.5 (http://www.repeatmasker.org) with RMBlast as the search engine and the following parameter settings: -s –no_is –

RepeatModeler using RepeatModeler version 4.0.5 (http://www.repeatmasker.org) library, the RepeatExplorer libraries (454 and Illumina) and the library from GenBank using BLASTN with an E value cut-off of 1 \( \times 10^{-10} \), against the non-redundant database using BLASTX with an E value cut-off of 1 \( \times 10^{-5} \), and submitted to Tandem Repeat Finder version 4.07b with default parameters 43.

RNA-seq reads from the five sequenced RNA samples were filtered for adaptors and quality trimmed, rRNA reads were identified and removed (ref. 44).

Assembled Trinity transcripts were mapped to the F. excelsior assembly using GMAP (r20141229) 51 and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed before downstream analyses (see below).

Identification of repetitive elements from the genome assembly. *De novo* identification of repetitive elements from the assembled ash genome sequence was conducted with RepeatModeler version 1.0.7 (http://www.repeatmasker.org/RepeatModeler.html) using RMBlast as the search engine. All unannotated (‘unknown’) repeat families from the RepeatModeler library were searched against a custom BLAST database of organellar genomes (see above) using BLASTN with an E value cutoff of 1 \( \times 10^{-10} \) and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed.

To prevent any captured gene fragments within repetitive element families causing the masking of protein coding genes within the ash assembly, the custom repeat libraries were pre-masked using the TAIR10 CDS data set (TAIR10_cds_20101214_updated; downloaded from http://www.arabidopsis.org). First, transposonPSI version 2 (http://transposonpsi.sourceforge.net) was run with the ‘nuc’ option to identify any transposable-element-related genes within the TAIR10 CDS data set. Sequences with a significant hit to transposable-element-related sequences (E value cut-off of 1 \( \times 10^{-5} \)) were removed from the TAIR10 CDS file \( (n = 308) \); a further 19 sequences that included the term ‘transposon’ in their annotation, but which did not have a hit using transposonPSI, were also removed. The filtered TAIR10 CDS data set was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and Illumina) and the library from RepeatMasker using RepeatMasker version 4.0.5 (http://www.repeatmasker.org) with RMBlast as the search engine and the following parameter settings: -s –no_is –

RepeatModeler using RepeatModeler version 4.0.5 (http://www.repeatmasker.org) library, the RepeatExplorer libraries (454 and Illumina) and the library from GenBank using BLASTN with an E value cut-off of 1 \( \times 10^{-10} \) and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed.

To prevent any captured gene fragments within repetitive element families causing the masking of protein coding genes within the ash assembly, the custom repeat libraries were pre-masked using the TAIR10 CDS data set (TAIR10_cds_20101214_updated; downloaded from http://www.arabidopsis.org). First, transposonPSI version 2 (http://transposonpsi.sourceforge.net) was run with the ‘nuc’ option to identify any transposable-element-related genes within the TAIR10 CDS data set. Sequences with a significant hit to transposable-element-related sequences (E value cut-off of 1 \( \times 10^{-5} \)) were removed from the TAIR10 CDS file \( (n = 308) \); a further 19 sequences that included the term ‘transposon’ in their annotation, but which did not have a hit using transposonPSI, were also removed. The filtered TAIR10 CDS data set was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and Illumina) and the library from RepeatMasker using RepeatMasker version 4.0.5 (http://www.repeatmasker.org) with RMBlast as the search engine and the following parameter settings: -s –no_is –

RepeatModeler using RepeatModeler version 4.0.5 (http://www.repeatmasker.org) library, the RepeatExplorer libraries (454 and Illumina) and the library from GenBank using BLASTN with an E value cut-off of 1 \( \times 10^{-10} \) and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed.

To prevent any captured gene fragments within repetitive element families causing the masking of protein coding genes within the ash assembly, the custom repeat libraries were pre-masked using the TAIR10 CDS data set (TAIR10_cds_20101214_updated; downloaded from http://www.arabidopsis.org). First, transposonPSI version 2 (http://transposonpsi.sourceforge.net) was run with the ‘nuc’ option to identify any transposable-element-related genes within the TAIR10 CDS data set. Sequences with a significant hit to transposable-element-related sequences (E value cut-off of 1 \( \times 10^{-5} \)) were removed from the TAIR10 CDS file \( (n = 308) \); a further 19 sequences that included the term ‘transposon’ in their annotation, but which did not have a hit using transposonPSI, were also removed. The filtered TAIR10 CDS data set was used to hard mask the RepeatModeler library, the RepeatExplorer libraries (454 and Illumina) and the library from RepeatMasker using RepeatMasker version 4.0.5 (http://www.repeatmasker.org) with RMBlast as the search engine and the following parameter settings: -s –no_is –

RepeatModeler using RepeatModeler version 4.0.5 (http://www.repeatmasker.org) library, the RepeatExplorer libraries (454 and Illumina) and the library from GenBank using BLASTN with an E value cut-off of 1 \( \times 10^{-10} \) and the DUST filter switched off. Any repeat families matching plastid or mitochondrial DNA were removed.
support with a Viridiplantae (without F. excelsior) protein database (≤50% BLAST high-scoring segment pair coverage) or where the CDS length was less than 100 bp (retaining those transcripts with ≥50% BLAST high-scoring segment pair coverage). Gene models were also excluded if they aligned with at least 30% similarity and 40% coverage to the TransposonPSI (version 08222010) library (http://transposonpsi.sourceforge.net/) and had at least 40% coverage by the RepeatModeler/RepeatMasker interspersed repeats. In addition, gene models that had at least 30% similarity and 60% coverage to the TransposonPSI library were excluded. Predicted protein-coding sequences were verified using the Fisher's exact test tool within the BLAST2GO plugin, using default parameters except for Annotation Cutoff = 55 and high-scoring segment pair-hit coverage cutoff = 40. Significantly enriched Gene Ontology terms were identified using the Fisher’s exact test tool within the plugin, where the reference set was the Gene Ontology terms for all genes, and a false discovery rate of 0.05 was used.

**Analysis of gene families.** The OrthoMCL pipeline (version 2.0.9) was used to identify clusters of orthologous and paralogous genes from F. excelsior and the following: Amborella trichopoda, Arabidopsis thaliana, barrel medic, bladdersword, coffee, grape, loblolly pine, monkey flower, poplar, and tomato (Supplementary Table 10). Input proteomes contained a single transcript per gene and were filtered with orthomcFilterFasta to remove any sequences of fewer than ten amino acids in length and/or >20% stop codons. Similar sequences were identified via an all versus all BLAST search for the 36,741 proteins remaining after filtering. The BLAST search was performed in the BLAST + package (version 2.2.29+), using an E value cut-off of 1 × 10^-5. BLAST results were filtered with orthomcPairs to retain protein pairs that match at least 50% of the length of the shorter sequence in the pair. Clustering of sequences was performed with mef9 (version 14.137) using a setting of 1.5 for the inflation parameter. The output from OrthoMCL was summarized using a custom Perl script to obtain counts of the number of sequences from each species belonging to each group. Venn diagrams for selected taxa were generated using InteractiVenn.

**European Diversity Panel sequencing.** DNA from the 37 European Diversity Panel trees was sequenced at the Earlham Institute on Illumina HiSeq, using paired-end insert sizes between 100 and 700 bp, and a read length of 150 bp. This generated an average of 63.6 million 150 bp reads (10.9× genome coverage) per tree. Filtering and trimming steps reduced this average to 55.3 million reads. An average of 85.8% of these reads per tree mapped to our reference genome. In addition, DNA reads from Danish Tree35 library ‘3077’ were downloaded from the Open Ash Dieback website (http://oadb.tsl.ac.uk/); these were 250 bp paired-end reads with an insert size between 200 and 400 bp. Tree35 is given the sample number 38PA in all further population analysis.

**SNP call validation using the KASP platform.** To test the reliability of SNP calls in the genome-wide SNP calling, we designed KASP assays for 53 SNPs, which ranged in their level of confidence (see Supplementary Table 13). None of the SNP calls tested by KASP were present in the reduced SNP set used for population genetic analyses. Primers were designed with a modified version of PolyMarker including FAM or HEX tails (FAM tail: 5′-GAAGTTGGAGCCAGTGTGCT-3′; HEX tail: 5′-GAAGTGGGAGGCAGCTAGTT-3′). The primer mix was prepared as recommended by the manufacturer (46 μl distilled H2O, 30 μl common primer (100 μM) and 12 μl of each tailed primer (100 μM)) (http://www.lgcgroup.com/ services/genotyping ). The reactions (2-μl mixes) were performed using the QuantStudio 5 (Life Technologies, Carlsbad, CA) according to the following protocol: hotstart at 95°C for 15 min, followed by ten touchdown cycles (95°C for 20s; touchdown 65°C, 1°C per cycle, 25 s) then followed by 30 cycles of amplification (95°C for 10s; 57°C for 60s). Fluorescence was detected on a Tecan Safire at ambient temperature. Genotypes were called using Stacks software (version 2.22.0.5; LGC Hoddesdon, UK). Four of the individuals did not amplify and were discarded from the analysis. The results of the calls are in Supplementary Data 7.

**European Diversity Panel genome-wide SNP calling.** The raw reads from the 37 trees in the European Diversity Panel (Supplementary Table 11) were aligned to the reference genome using Bowtie 2.2.5 (ref. 81). The alignments were converted to BAM format and duplicated reads were removed with samtools 1.2 (ref. 82). To assign each read to its corresponding tree, the flag ‘rg’ was added to each BAM file with picard tools 1.119 (http://broadinstitute.github.io/picard/). SNPs were called with freebayes 1.0.2 (ref. 10) to produce a VCF file. The SNPs with quality less than 300 were filtered with bio-samtools 2.1 (ref. 83). SnEff 4.1g (ref. 84) was used to predict the effect of the putative SNPs (see Supplementary Table 12). Genic regions were within 5kb from a gene model. Amino-acid changes were labelled as missense, variant.
of 50bp, and were also trimmed of any adaptor and repetitive telomere sequences. Filtered reads were mapped to the reference assembly using the 'Map Reads to Reference' tool in the CLC Genomics Workbench, setting both similarity match and length match parameters to 0.95. Regions with coverage of between 5 and 30 reads in all samples were extracted using the 'Create Mapping Graph', 'Identify Graph Threshold Areas' and 'Calculate Track' tools. These extracted regions totalled 20.6Mbp (2.3% of the genome).

Variant calling was performed on a read mapping pooled from all samples, using the 'Low Coverage Variant Caller' tool in the CLC Genomics Workbench, with the coverage-restricted regions from the previous step used as a track of target regions. This prevented variants being called where some samples did not have read coverage, and in the organellar scaffolds where the read coverage was very high. The following parameters were changed from default: Ignore positions with coverage above = 1,000, Ignore broken pairs = no, Ignore non-specific matches = Reads, Minimum Coverage = 190 (38 samples with at least 5 reads each should have a combined total coverage of > 189), Minimum Count = 10, Minimum Frequency = 5%, Base Quality Filter = Yes, Neighbourhood radius = 5, Minimum Central Quality = 20, Minimum neighbourhood quality = 15, Read Direction Filter = yes, Direction Frequency = 5%. As a result, 529,812 variants were called, comprising 468,237 SNPs, 14,850 equal replacements (where >1 nucleotide is replaced by an equal number of nucleotides), 26,043 deletions, 19,085 insertions and 1,597 unequal replacements (where at least one SNP lies directly beside an indel). The average quality of all reads at these variant positions was 36.2.

To genotype each sample individually at the variant loci called in the previous step, the 'Identify Known Mutations from sample mappings' tool within the CLC Biomedical Genomics workbench was used. The workbench takes a track of known variants as input (such as those called from the pooled read mapping) and reports the presence, absence, coverage, count and other statistics of each variant locus in the given mapping using the read mapping of another sample (in this case, the read mapping from each of the 38 trees). The 'Identify Candidate Variants' tool was then used to filter variants with a minimum coverage of 5, minimum count of 3 and minimum frequency of 20%. VCF files for each tree were exported from the CLC Workbench and merged into one file using the vcf-merge tool from VCFTools. The merged VCF file was then filtered using vcf tools, to remove indels, multi-allelic loci, and loci with a minimum allele frequency <0.05, with 394,885 SNP loci remaining. This high-quality SNPs with comprehensive knowledge of the genotype of every sample was referred to as the ‘reduced SNP set’ and used for further population analyses.

To visualize similarities and differences among the genomes of the European Diversity Panel, PCA was performed using the SNPRelate version 1.4.2 (ref. 87) package in R version 3.1.2. The filtered VCF file was converted into gds using the snpgdsVCF2GDS command, and was filtered on a linkage disequilibrium value of 0.1 using the snpgdsLDpruning command, leaving 34,607 SNPs. PCA was performed on the pruned set of SNPs using the snpgdsPCA command with default options, and the results of the first three PCs were plotted in R.

To analyse population structure in the European Diversity Panel, scaffolds were selected that contained 10 or more SNPs in the filtered VCF file (8,955 nucleotides in total). Three different sets of SNPs were selected at random from each of these scaffolds, and placed into three different files in STRUCTURE input format (26,865 SNPs in total, 8,955 in each set). STRUCTURE version 2.3.4 (ref. 88) was run with admixture from k = 1 to k = 20 for each of the three sets of SNPs, with both BURNIN and NUMREPS set to 100,000. All output results were run through Structure Harvester Web version 0.6.94 (ref. 89), which found k = 3 to have the largest Δk value of 32.91 (Extended Data Fig. 3). Next, the three runs of k = 3 were used as input into CLUMP version 1.1.2 (ref. 90) to align the clusters, and samples within each cluster. Aligned results were imported back into STRUCTURE version 2.3.4 to generate Q value bar plots. Average Q values from the three runs were used to generate a map with pie charts, using Tableau version 9.3 (Tableau, Seattle, USA) with Tableau base-map country outlines. Each section of the pie represented a value bar plots. Average values from the three runs were used to generate a map with pie charts, using Tableau version 9.3 (Tableau, Seattle, USA) with Tableau base-map country outlines. Each section of the pie represented

Effective population size estimation by linkage disequilibrium in the European Diversity Panel was performed using the program SNPRelate version 1.1 (ref. 92), which takes genome-wide polymorphism data from several individuals in a population as input. The European Diversity Panel filtered VCF file with the reduced SNP set of 38 trees (the same as used in PCA and STRUCTURE analysis) was converted into Map and Ped files. The third column in the Map file (linkage distance in morgans) was set to zero for all SNPs, as these values were unknown and SNPRelate calculates this value from each SNP's physical distance. SNPRelate was then run with a minimum distance between SNPs of 10,000 bp and a maximum of 400,000 bp, with Sved's modifier for recombination rate, and with 50 bins. Estimated effective population sizes were plotted in R (Extended Data Fig. 3c), as well as linkage disequilibrium decay over distance between 100 and 300,000 bp (Fig. 2e).

Simple-sequence repeat analysis. To develop accessible population genetic markers, the repeat masked version 0.4 2451S genome was mined for simple sequence repeat (SSR) sequences (a repeat motif of 2–5bp in length repeated a minimum of five times) using the QDD version 3.1 pipeline. Downstream QDD version 3.1 pipelines screened SSR loci (inclusive of the SSR repeat motif and 200bp forward and reverse flanking regions) for singleton loci in an all-against-all BLAST (–task blastn –evalue 1e-40 -lcase_masking -soft_masking true) and designed primer pairs within 200bp flanking regions using PRIMER3 software. The approximately 31,300 singleton SSR loci identified in the ash genome were screened using RepeatMasker Open-4.0 (http://www.repeatmasker.org) in QDD version 3.1 to eliminate loci that hit known transposable elements in the RepBase Viridiplantae repeat library (http://www.girinst.org), leaving about 28,800 SSR loci. The final primer table output by the QDD version 3.1 pipeline allows selection of the best primer pair design for each SSR loci. To select candidate markers for further development, these primer pairs were filtered according to parameters provided by QDD version 3.1. The selected SSR loci had a: maximum primer alignment score of 5; minimum 20bp forward and reverse flanking region between SSR and primer sequences; high-quality primer design (defined by QDD pipeline as an absence of homopolymer, nanosatellite and microsatellite sequence in primer and flanking sequences); and minimum number of 7 motif repeats within the SSR sequence. This filter was applied to 857 SSR loci, which were screened against the combined custom ash repeat library for v0.5 of the 2451S genome assembly (see above: Analysis of repetitive DNA) via a BLASTN search with an E value of 1 × 10−10 in the BLAST+ package (version 2.2.31-1). Elimination of all sequences with a hit to known repetitive elements left 681 candidate loci. These were compared with the v0.5 assembly via a BLASTN search with an E value cut-off of 1 × 10−10. This returned a set of 664 loci with a unique match to the v0.5 assembly for use as population genetic markers (see Supplementary Data 1).

In silico analysis of allelic diversity (that is, locus polymorphism) of these SSR loci was performed by screening a subset of loci (366) against a variance table (containing all locations corresponding for the European Diversity Panel. Approximately half (48%) of the loci tested were variable among 37 of the sequenced genomes (sample 38 not included). Twenty candidate SSR loci with the greatest in silico allelic diversity were selected for wet laboratory testing on seven individuals from the European Diversity Panel. Primer pairs with a fluorescent tag on the 5′ end of the forward primer (FAM, HEX or TAM) were used. For singleplex PCR, primer aliquots were used at a concentration of 10pmol/μL. Multiplex PCR amplification of target regions was performed in singleplex reactions with a final reaction volume of 10μL containing 1μL genomic DNA, 0.2μL of each primer (10pmol/μL), 3.6μL of RNase free water, and 5μL of Qagen Type-it Multiplex PCR Master Mix. The amplification conditions were as follows: 5 min at 95°C, 18 cycles of 30s at 95°C, 90s at 62°C with a 0.5°C reduction per cycle, 30 s at 72°C, 20 cycles of 30s at 95°C, 1 min 30s at 51°C, 30 s at 72°C; a final extension step of 30 min at 60°C. PCR products were diluted to 1:10 with distilled H2O and run on (an Applied Biosystems 3730xl 96 capillary sequencing instrument with Applied Biosystems GeneScan 400HD Rox
Two of the 20 significantly associated GEMS in the present study, FRAEX38873_v2_000048360.1 (P = 1.77 × 10⁻⁶) and FRAEX38873_v2_000048340.1 (P = 3.48 × 10⁻⁷), did not have high BLAST similarity to GEMS found in the previous study. However, these GEMS were highly similar to a cDNA transcript containing a predictive A/G SNP (termed a cSNP) identified previously, where presence of a G allele was associated with low damage scores. Both of these GEMS contained the ‘less susceptible’ G variant. A third paralogous gene in this family with the A variant was also found (FRAEX38873_v2_000184430.1), and was not included in the assay with damage score analysis due to low RPKM. The present study therefore resolves this cSNP marker into three paralogous genes, two fixed for a ‘less susceptible’ G nucleotide, and one a ‘susceptible’ A nucleotide.

These five GEMS were used as qRT–PCR, and, in the case of FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000048340.1 RT–PCR, to a small test panel of 58 Danish accessions (henceforth Danish Test Panel) to assess their predictive capabilities in a similar way as in ref. 3. Unlike this previous study, however, ratios between the bases of the FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000048340.1 were scored by eye (instead of simply scoring the presence or absence of the ‘less susceptible’ nucleotide), to estimate levels of gene expression for the ‘less susceptible’ paralogue, while maintaining the simplicity of the assay. These ratios and the qRT–PCR assays for the other three GEMS were combined into a single predicted damage score for each of the Danish Test Panel, which could then be compared with the observed damage scores for these trees. The combined prediction was correlated with the log mean damage scores for 2013–2014 (r² = 0.25, P = 6.9 × 10⁻⁵) which gave a small improvement in predictive power from the previous analysis (r² = 0.24, P = 8.4 × 10⁻⁸).

Screening of UK F. excelsior accessions for markers of reduced susceptibility to ADB. Four markers were selected for predictive marker assays on the basis of this analysis and previous work on the Danish Test Panel of 58 trees. The three GEM markers most highly associated with disease damage were assayed by qRT–PCR using the following primer combinations: FRAEX38873_v2_000261470.1 (GTCCGAGGAGATGTCAGTCAT, ATCTTGTGGGACCTATCG), FRAEX38873_v2_000199610.1 (GGTGGAGGAGAAAGTGCAATG, TGGCCTTTGAGAAGAACA), FRAEX38873_v2_000173540.1 (AGGGCAACGTTGAAACAT, TAGGGCTTTTCTTAGCCTGTTGAC) and GAPDH reference (CTGGGAGTCGCTTCAAGAAG, CGTCAAACTAACACCAGGAAA).

Using RNA extracted from the British Scoring Panel, qRT–PCR reactions were performed with SYBR Green fluorescence detection in a qPCR thermal cycler (ViiA 7, Applied Biosystems, San Francisco, California) using optical grade 384-well plates, allowing all reactions to be performed simultaneously for each target gene. Each reaction was prepared using 3 μl from a 2 ng/μl dilution of cDNA derived from the RT reaction, 5 μl of SYBR Green PCR Master Mix (Applied Biosystems), 200 nM forward and reverse primers, in a total volume of 10 μl. The cycling conditions were 2 min at 50 °C, 10 min at 95 °C, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min with the final dissociation at 95 °C for 15 s, 60 °C for 1 min and 95 °C for 15 s. Three technical replicates were used for quantification analysis. Melting curve analysis was performed to evaluate the presence of primer and amplicons. The specificity of the primers and the amplicons were verified by amplicon sequencing (GATC Biotech LIGHTTru). The results were exported as raw data, and the LinRegPCR99 software was used for baseline correction. The resulting means of triplicate N0 values, representing initial concentrations of a target and reference gene, were used to analyse gene expression. For each marker, the set of qRT–PCR quantities were standardized and rescaled to better emulate the range of RPKM values observed in the original association panel, and then predicted damage scores generated using the regression coefficient and constant from the GEM associations. An additional GEM marker was assayed as a cSNP by PCR using 1 μl of DNA. The 11.5 μl thermal cycling conditions were: 1× HotStart ThermoPol Plus Buffer (2X), 200 nM forward (GTGGTCTCTGTGCAGCCG) and reverse (TCCATGATCCACG) primers in a total volume of 25 μl. The touchdown PCR was performed in using a BIORAD Tetrad PCR machine with the following cycling conditions: 5 min at 94 °C, followed by 15 cycles of 94 °C for 30 s, 63 °C for 30 s – 1 °C per cycle, 72 °C for 1 min, and 30 cycles of 94 °C for 30 s, 53 °C for 30 s, 72 °C for 1 min and a final elongation step at 72 °C for 7 min. Sangre sequences obtained using the forward primer co-amplify GEM FRAEX38873_v2_000048360.1, which is highly associated with ADB disease damage, and another member of the gene family that is not. Owing to a polymorphism in the non-specific primer products and primer dimers. The specificity and uniqueness of the primers and the amplicons were verified by amplicon sequencing (GATC Biotech LIGHTTru). The results were exported as raw data, and the LinRegPCR99 software was used for baseline correction. The resulting means of triplicate N0 values, representing initial concentrations of a target and reference gene, were used to analyse gene expression. For each marker, the set of qRT–PCR quantities were standardized and rescaled to better emulate the range of RPKM values observed in the original association panel, and then predicted damage scores generated using the regression coefficient and constant from the GEM associations.
analysis. G.A peak height ratios were approximated directly from the sequence chromatograms using Softgenetics Mutation Surveyor software for the British Screening Panel and the Danish Test Panel. These ratios were then standardized and rescaled to the RPKM values for FRAEX38873_v2_000048360.1 to predict damage scores as before.

Combined predictions were made by ranking and standardizing the individual predictions for all four markers, and then calculating the mean rank score for each individual tree (Supplementary Data 6). Combined predictions were calculated for the Danish and the British Panel, compared with the observed ADB damage scores to ensure that the assay was predictive (Fig. 3).

The four assays were applied in the same way to analyse a panel of 130 accessions originating from across the UK range of *F. excelsior* (*British Screening Panel*). Strikingly, when assayed by RT–PCR, expression of the G′ variant paralogues was seen at much higher frequency in the British Screening Panel than in the Danish panels and the mean G:A ratio across the British Screening Panel was 0.67 compared with a mean of 0.03 observed in the Danish Test Panel. Likewise, the gene expression estimates for the British Screening Panel exhibited wider ranges and were more favourable in terms of their expected effect on damage scores. The qRT–PCR results for the GEMs negatively correlated with disease damage (FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) exhibited higher mean expression in the UK (0.01 ± 0.01 and 0.12 ± 0.14) versus the Danish Test Panel (0.09 ± 0.08, 0.12 ± 0.11), and the positively correlated FRAEX38873_v2_000173540.4 was on average expressed at a lower level in the British Screening Panel (0.48 ± 0.26) than the Danish Test Panel (0.59 ± 0.17). As expected, this translated to lower combined predictions for ADB damage in the British Screening Panel. Only 9% of the Danish Test Panel accessions were predicted to have a low damage score (defined as 25% canopy damage or less) compared with 25% of the British Screening Panel (Fig. 3).

**Analysis of predicted gene trees.** To predict the susceptibility of the reference tree 24515 to ADB, we calculated RPKM values for the five GEM marker CDS models (FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1, FRAEX38873_v2_000048360.1, FRAEX38873_v2_000261470.1 and FRAEX38873_v2_000199610.1) from leaf transcriptome read data. We also did this for each of the trees in the Danish Scoring Panel, and the average of these predictions was taken to provide combined predictions. The top and bottom quartiles from the distribution of predicted scores, which represent the trees with the most susceptible and least susceptible gene expression patterns at these five loci, were then correlated with the RPKM values for the genome sequenced tree 24515 (Extended Data Fig. 4).

RPKM data were also generated for four tissue types: leaf, flower, cambium and root, of the parent of sequenced tree 24515 by mapping raw reads to the CDS reference as before. RPKM data for the 20 CDS models found to be significantly associated with susceptibility to ADB in the GEM analysis were selected and compared for the four tissue types.

The five CDS models represented in the ADB susceptibility predictions were translated using the standard codon usage table and were searched against the non-redundant database in GenBank using BLASTp with default settings to identify top hits to protein sequences in *A. thaliana*; FRAEX38873_v2_000199610.1 and FRAEX38873_v2_000048360.1 show high-quality analysis to AGAMOUS-LIKE 42/FOREVER YOUNG FLOWER (AGL42/FYY; AT3G62165); FRAEX38873_v2_000173540.4, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 have top hits to SHORT VEGETATIVE PHASE/AGL71 (SVP/AGL22-like) (Extended Data Fig. 5). These all belong to the SVP/StMADS11 group of type II MADS-box genes. To find potential orthologues of these proteins, we further expanded these data sets to include sequences from the OrthoMCL clusters containing *A. thaliana* proteins from closely related MADS lineages, as identified by previous phylogenetic analyses of type II MADS-box sequences.

Preliminary phylogenetic analysis of these data sets revealed that, despite showing high sequence similarity in BLAST searches, FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 do not fall within the clade containing SVP/AGL22 and similar *A. thaliana* sequences. Therefore, to identify potentially more closely related sequences we performed a BLASTp search of FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 against the complete set of MADS protein sequences in the Arabidopsis genome (see Supplementary Table 10), using the BLAST+ package (version 2.2.31) with an *E* value cut-off of 1 × 10⁻⁵. The BLAST+ results contained a number of putative transposable-element-related genes during annotation. This identified several highly similar sequences from other species with better ranking BLAST hits than those to the *A. thaliana* proteins. These sequences belong to a single OrthoMCL cluster, and include a tomato (*S._lycopersicum*) sequence from the apparent orthologue of the potato (*S. tuberosum*) SSMADS1 gene; all sequences from this cluster were added to the SVP/AGL22-like data set, along with the potato SSMADS1 protein (GenBank accession number ACH53556.1). Sequences for both data sets were aligned using M-Coffee, followed by the T-Coffee web server (http://www.tcoffee.org; last accessed 7 December 2016) with the following parameter settings: Mpcma_msa Mmafft_msa Mcluswalt_msa Mmproteins Mpsa Mmpcons_msa Mmci coffee_msa output = score_html clustalw_align fastaln_score ascii phylib -tree -maxtrees = 150 -maxseq = 2500 -case = upper -seqnos = on -outerdir = input -run_name = result -multicore = 4 -quiet = stdout. Positions in the alignments containing consensuses scores of <6 from M-Coffee were removed; filtered alignments were then rerun through the TCS tool with the T-Coffee web server and any positions with a reliability score of <6 were removed. Recombination was tested for in the filtered alignments using GARD. Analyses were run via the Datamonkey server (http://www.datamonkey.org; last accessed 1 June 2016) under the best-fit model of evolution (selected with the corrected Akaike’s information criterion) with 3杰 rate variation and three rate classes. No breaks with significant topological incongruence at P ≤ 0.05 were detected for either data set. Phylogenetic analysis of each data set was conducted using Bayesian inference in MrBayes and maximum likelihood in RAxML; input alignments are provided in Supplementary Data 8. MrBayes (version 3.2.5 (ref. 104)) was run using the mixed amino acid model, to allow models of protein sequence evolution to be fit automatically across the alignments; the following parameter settings were used for each data set: preset aamodeltype = mixed, mcmcnruns = 2, nchains = 4, ngen = 100000, samplefreq = 1000. Parameter values from both runs for each data set were viewed in TRACER version 1.6 (http://beast.bio.ed.ac.uk/Tracer) to confirm that effective sample sizes of > 200 had been obtained for each parameter and stationarity reached. Trees sampled during the first 100 000 generations of each run were discarded as the burn-in; trees and parameter values were summarized in MrBayes using the sum and sump commands. RAxML (version 8.2.8 (ref. 105)) was run using the option to automatically determine the best protein substitution model, with 1000 replicates of the rapid bootstrap algorithm; parameter settings were as follows: raxmlHPC -f a -x 13102 -p 29503 -n 1000 -m PROTGAMMAAUTO. The phylogenetic analysis suggested that FRAEX38873_v2_000173540.4 is a likely orthologue of the *A. thaliana* SVP/AGL22 gene, or possibly AGL24, whereas FRAEX38873_v2_000048340.1 and FRAEX38873_v2_000048360.1 appear orthologous to the potato SSMADS1 gene (Extended Data Fig. 5). These all belong to the SVP/SSMADS11 group of type II MADS-box genes. FRAEX38873_v2_000048360.1 and FRAEX38873_v2_000199610.1 cluster with the *A. thaliana* SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1)-like proteins AGL42, AGL71 and AGL72 (Extended Data Fig. 5). The two other major clades within the phylogenetic tree include the AGL20/SOCI protein and the AGL14 and AGL19 proteins (Extended Data Fig. 5); together, the AGL42/AGL71/AGL72, AGL20- and AGL14/AGL19-containing clades are known as the SOC1/TM3 group of type II MADS-box proteins.

In *A. thaliana*, *AGL20- and AGL14-encoding* genes have redundant functions in controlling flowering time and appear to be regulated by both AGAMOUS and AGAMOUS-LIKE 71 (SVP/AGL22, AGL20/AGL14) and SVP/AGL22-like proteins. These sequences belong to a SOC1-like protein clade with wider functions.

To evaluate the potential of these clades to contribute to tolerance or resistance, we compared triplicate samples from five low-susceptibility Danish trees (R-14164C, R-14184A, R-14193A, R-14198B, R-14181) and five high-susceptibility trees (R-14169, R-14127, R-14156 R-14120, 25UTaps).
Three leaflets from each triplicate sample were freeze-dried and gently crushed to mix tissue. Approximately 100–150 mg was ground to a fine powder using a TissueLyser (Qiagen), and 10 mg was extracted in 400 μl 80% MeOH containing d5-IAA internal standard at 2.5 ng/ml [1H2]indole-3-acetic acid; OIChemIm, Czech Republic), centrifuged (10,000 g, 4 °C, 10 min) and the pellet re-extracted in 80% MeOH. The pooled supernatants were filtered through a 0.2 μm syringe filter (Pall Corporation, UK).

These leaf extracts (5 μl) were analysed using a Polaris C18 1.8 μm, 2.1 mm × 250 mm microbore reversed-phase analytical column (Agilent Technologies, Palo Alto, California, USA) and samples resolved on an Agilent 1200 series Rapid Resolution HPLC system coupled to a quadrupole time-of-flight QToF 6520 mass spectrometer (Agilent Technologies, Palo Alto, California, USA). Buffers were as follows: positive ion mode; mobile phase A (5% acetonitrile with 0.1% formic acid), mobile phase B (95% acetonitrile with 0.1% formic acid); negative ion mode; mobile phase A (5% acetonitrile with 1 mM ammonium fluoride), mobile phase B (95% acetonitrile). The following gradient was used: 0–10 min, 0% B; 10–30 min, 0–100% B; 30–40 min, 100% B. The flow rate was 0.25 ml/min and the column temperature was held at 35 °C throughout. The source conditions for electrospray ionization were as follows: gas temperature was 325 °C with a drying gas flow rate of 91/min and a nebulizer pressure of 35 pounds per square inch gauge. The capillary voltage was 3.5kV in both positive and negative ion mode. The fragmentor voltage was 115 V and skimmer 70 V. Scanning was performed using the autoMS/MS function at four scans per second for precursor ion surveying and three scans per second for MS/MS with a sloped collision energy of 3.5 V per 100 Da with an offset of 5 V.

Positive and negative ion data were converted into mzData using the export option in Agilent MassHunter. Peak identification and alignment was performed using the Bioconductor R package xcms100 and features were detected using the centroid method108 for high-resolution liquid chromatography/mass spectrometry data in centroid mode at 30 p.p.m. Changes from the default parameters were mzdff = 0.01, peakwidth = 10–80, noise = 1000, prefilter = 3,500. Peaks were matched across samples using the density method with a bw = 5 and mzdff = 0.025 and retention time correlated using the obwp algorithm with protStep = 0.5. Missing peak data were filled in the peaklists generated from the ADB low-susceptibility ash leaf samples compared with the peaklists generated from the ADB susceptible leaves. The resulting peaklists were annotated using the Bioconductor R package, CAMERA109. The peaks were grouped using 0.05% of the width of the full width at half maximum, and groups correlated using a P value of 0.05 and calculating correlation inside and across samples. Isotopes and adducts were annotated using a 10 p.p.m. error.

Statistical analysis and modelling was performed using MetaboAnalyst version 3.0 with the following parameters. Missing values were replaced using a KNN imputing missing value estimation. Data were filtered (40%) to remove non-informative variables using the interquartile range. Samples were normalized using the internal standard d5-IAA (POS: M181T1448; NEG: M179T1382). Data were auto-scaled. Peaks from the three replicates were aligned with xcms for both positive and negative mode and features tested for practical significance to determine the differences between the tolerant and susceptible genotypes. In addition, PLS-DA was performed using MetaboAnalyst, allowing the discrimination of tolerant and susceptible genotypes on the basis of their metabolic profiles (Fig. 4a).

The individual features (putative metabolites) that contributed to the separation between the different classes were further characterized. We first applied a range of univariate and multivariate statistical tests to determine the importance of these features. This included variable influence on the projection (VIP) values derived from PLS-DA scores, practical significance, t-test, P value, Benjamini and Hochberg false discovery rate P value, effect size and Random Forest analysis, and MS/MS fragmentation network analysis. For example, using Random Forest, significant features were ranked by mean decrease in classification accuracy with 14 out of 15 susceptible samples (out-of-bag error: 0.033; class error 0.07) and 15 out of 15 tolerant samples correctly classified.

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

Identification was not possible for those features with no fragmentation, or lacking significant supporting adducts. Many features of interest were identified but required further work to provide confident attributions, while some features did not provide fragmentation patterns. We thus restricted further identification and characterization to a highly discriminatory class of compounds of the iridoid glycosides and predominantly compounds previously recorded in Oleaceae, summarized in Extended Data Figs 6–9 and Supplementary Data 9. We validated these identifications using three methods: MS/MS fragmentation networking (Fig. 4c), MS/MS mirror plot (Extended Data Fig. 6) and accurate mass MS/MS product ion structure correlation (Extended Data Fig. 7). The MS/MS fragmentation network was generated after extracting the m/z values of the MS/MS product ions from the discriminatory features using MassHunter Qualitative Analysis Version 4 and visualized using Cytoscape, indicating product ion masses that had been previously reported from fragmentation of iridoid glycosides110.

Further validation was performed through a mirror plot comparing the MS/MS spectra of four features (N2–N1) detected in negative mode with an electrospray ionization-time of flight/ ion trap-mass spectrometry (ESI-TOF/IT-MS) spectrum of elenolic acid glucoside taken from the literature111. Finally, the accurate masses of MS/MS product ions from four discriminatory features identified in negative mode (N2–N1) were correlated with the structure of the putatively identified compound using MassHunter Molecular Structure Correlator (Agilent).

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

URL. Genome website: http://www.ashgenome.org.

Data availability. The reference tree is growing at Earth Trust with accession number 24515. Transcriptome and DNA reads and the final assembly for the 24515 genome sequence, as well as RNA reads for parent tree and raw reads and consensus read mappings of the European diversity panel trees, have been deposited in European Nucleotide Archive under project accession code PRJEB4958 (http://www.ebi.ac.uk/ena/data/view/PRJEB4958). Metabolomic data that support the findings of this study have been deposited in MetaboLights under accession code MTBLS372 (http://www.ebi.ac.uk/metabolights/MTBLS372). All other data are available from the corresponding author upon reasonable request.
Extended Data Figure 1 | Completeness and coherence of annotation models. a, Assessment of transcript completeness for the *F. excelsior* gene set. Transcripts were classified as full-length, 3′-end, 5′-end, internal, coding (open reading frame predicted but no BLAST support), unknown (no BLAST support), mis-assembled and putative ncRNA using Full-lengtherNEXT (version 0.0.8); 76.43% of transcript models were identified as complete. b, Coherence in gene length between *F. excelsior* and *M. guttatus* proteins. BLAST analysis ($1 \times 10^{-5}$) identified 2,576 proteins that had reciprocal best hits to 2,605 *M. guttatus* proteins identified as single copy in *M. guttatus*, *S. lycopersicum*, *S. tuberosum* and *V. vinifera* (Phytozome). A high coherence in gene length was found between *F. excelsior* and *M. guttatus*: $r > 0.917$. 

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Extended Data Figure 2 | Synteny between ash and monkey flower. Syntenic dotplot between ash (vertical axis) and monkey flower (horizontal axis) showing regions of multiple synteny. Scaffolds equal to approximately 75% of the ash genome assembly for which syntenic blocks were not detected are not shown. For clarity, small scaffold names are omitted.
Extended Data Figure 3 | Population structure of *F. excelsior* in Europe.

**a**, Results from STRUCTURE; three replicates were run for \( k = 3 \), with each replicate using a different set of 8,955 SNPs as input. Numbers refer to samples, whose locations are given in Supplementary Table 11.

**b**, \( \Delta k \) values for three runs of STRUCTURE of each value of \( k \) between \( k = 2 \) and \( k = 19 \); \( k = 3 \) has the highest \( \Delta k \) value of 32.91.

**c**, Effective population size history estimated using the SNeP program, with genotype information from all 38 diversity panel samples at 394,885 SNP loci.
Extended Data Figure 4 | Prediction of susceptibility of reference tree. RPKM values for leaf material from the low heterozygosity reference tree 2451S for five CDS models predictive for ADB. These are shown next to expression profiles for the Danish Scoring Panel with the least susceptible and most susceptible expression patterns according to the GEM analysis.
Extended Data Figure 5 | Investigation of the function of GEMs for low susceptibility to ADB. Unrooted maximum likelihood trees from the RAxML analyses. a, Best scoring maximum likelihood tree from the phylogenetic analysis of SVP/AGL22 and SIMADS11-like sequences. Nodes with bootstrap support values of at least 70 from the maximum likelihood analysis and posterior probabilities of at least 0.95 from the Bayesian analysis are indicated with asterisks. F. excelsior sequences are shown in blue; A. thaliana sequences in red. Four-letter taxon codes at the start of sequence names, where present, follow those in Extended Data Table 1. Sequence names are those from the original data files used for the orthoMCL analysis (see Supplementary Table 10), with the exception of the StMADS11 protein from potato, where the GenBank accession number is given. Common names for selected genes/proteins are annotated on the trees. Scale bars indicate the mean number of substitutions per site.
Extended Data Figure 6  MS/MS mirror plot of elenolic acid glucoside (ESI-TOF/IT-MS) compared with four negative mode features (N2, N3, N4 and N5). The spectra share four product ions in common: m/z 179, 223, 371 and 403 (elenolic acid glucoside molecular ion). These product ions correspond to a loss of a methyl and hydroxyl group (403–371), loss of hexose (403–223), which is followed by a loss of CO2 (223–179). Elenolic acid corresponds to the secoiridoid part of oleuropein-related compounds, suggesting that these four compounds are secoiridoids112.
Extended Data Figure 7 | Identification of MS/MS product ions for four iridoid-glycoside-related features observed in negative mode. Predicted structure for key m/z peaks using Molecular Structure Correlator (Agilent) and the structure of putative identities. Bonds and atoms in black are present in that product ion, whereas grey indicates loss.
Extended Data Figure 8 | Identification of iridoid-glycoside-related metabolites in positive mode. Box-plots showing abundance (log₂ transformed) of features in positive mode discriminating between five different genotypes of high- (TOL) and low- (SUS) susceptibility ash trees.
Extended Data Figure 9 | Identification of metabolites. MS/MS fragmentation product ion data of features discriminating between five different genotypes of high- (TOL) and low- (SUS) susceptibility ash trees in positive mode. Corresponding box-plots are presented in Extended Data Fig. 8.
Extended Data Table 1 | The 20 largest clusters in *F. excelsior* from the OrthoMCL analysis of 11 species showing the number of sequences from each species belonging to the clusters

| OrthoMCL cluster name | Putative gene family name(s)/function(s) | FEXC | ATHA | ATRI | CCAN | MGUT | MTRU | PITA | PTRI | SLYC | UGIB | VVIN |
|-----------------------|----------------------------------------|------|------|------|------|------|------|------|------|------|------|------|
| OG_00001              | Pentatricopeptide repeat (PPR) superfamily, Tetrapartite repeat (TPR)-like superfamily | 102  | 91   | 35   | 101  | 103  | 107  | 212  | 105  | 93   | 73   | 118  |
| OG_00003              | Leucine-rich repeat receptor-like protein kinase family, C/EBP-like transcription factors/leucine zipper domains | 81   | 40   | 34   | 112  | 52   | 112  | 121  | 114  | 50   | 24   | 63   |
| OG_00005              | Subtilase family, Subtilisin-like serine endopeptidase family protein/identical protein binding, serine-type endopeptidase activity | 63   | 46   | 42   | 50   | 95   | 88   | 49   | 67   | 71   | 21   | 65   |
| OG_00006              | S-box leucine-rich repeat protein kinase family, Putative receptor-like serine/threonine protein kinases/protein amino acid phosphorylation, recognition of pollen. | 58   | 32   | 7    | 42   | 43   | 125  | 9    | 183  | 53   | 1    | 52   |
| OG_00007              | Leucine-rich repeat protein kinase family, HIT-type Zinc finger family protein/serine/threonine kinase activity, kinase activity, ATP binding protein | 55   | 8    | 7    | 161  | 64   | 77   | 59   | 47   | 41   | 12   | 26   |
| OG_00012              | Laccase family | 43   | 18   | 14   | 23   | 20   | 23   | 54   | 54   | 27   | 8    | 43   |
| OG_00019              | Calcium-dependent protein kinase family | 40   | 31   | 11   | 16   | 23   | 25   | 9    | 28   | 28   | 24   | 16   |
| OG_00039              | Wall-associated kinase family | 40   | 19   | 3    | 8    | 34   | 20   | 6    | 46   | 9    | 0    | 10   |
| OG_00010              | Major facilitator superfamily | 39   | 22   | 20   | 25   | 28   | 49   | 54   | 40   | 22   | 20   | 25   |
| OG_00015              | P-glycoprotein family/ATPase activity, coupled to transmembrane movement of substances. | 37   | 22   | 14   | 25   | 23   | 39   | 38   | 36   | 22   | 10   | 20   |
| OG_00021              | n/a | 34   | 0    | 0    | 167  | 18   | 1    | 0    | 23   | 3    | 0    | 0    |
| OG_00037              | Cellulose synthase family (CESA), Cellulose synthase-like proteins/cells wall biosynthesis. | 31   | 16   | 14   | 12   | 14   | 21   | 16   | 28   | 17   | 19   | 16   |
| OG_00004              | LRR and NB-ARC, and NB-ARC domain-containing disease resistance proteins/ATP binding protein | 30   | 2    | 0    | 264  | 44   | 206  | 3    | 115  | 15   | 2    | 81   |
| OG_00028              | Cytochrome P450, family 7, subfamily B | 29   | 28   | 13   | 35   | 17   | 47   | 0    | 26   | 21   | 0    | 5    |
| OG_00026              | FAD-binding Berberine family | 28   | 27   | 4    | 27   | 26   | 28   | 1    | 63   | 19   | 5    | 4    |
| OG_00022              | Putative ligand-gated ion channel subunit family | 28   | 20   | 24   | 18   | 37   | 8    | 24   | 43   | 11   | 11   | 21   |
| OG_00025              | M航uka-like receptor-like protein kinase family | 27   | 17   | 9    | 25   | 31   | 43   | 1    | 41   | 19   | 12   | 9    |
| OG_00016              | Pleiotropic drug resistance family, ABC-2 and Plant PDR ABC-type transporter family/ nucleoside-triphosphate ATPase activity, nucleotide binding, ATP binding | 27   | 16   | 15   | 23   | 20   | 33   | 43   | 29   | 25   | 13   | 33   |
| OG_00059              | Leucine-rich repeat protein kinase family, Plasma membrane LRR receptor-like serine/threonine kinase proteins, Somatic embryogenesis receptor-like kinase proteins/protein serine/threonine kinase activity, kinase activity, ATP binding. | 28   | 14   | 7    | 8    | 14   | 15   | 7    | 20   | 13   | 11   | 11   |
| OG_00085              | Raffinose synthase family | 26   | 5    | 12   | 9    | 13   | 12   | 9    | 13   | 6    | 12   | 10   |

Clusters containing at least five more sequences from *F. excelsior* than for the other asterid species (underlined) are shown in bold. FEXC, *F. excelsior*; ATHA, *A. thaliana*; ATRI, *A. trichopoda*; CCAN, *C. canephora*; MGUT, *M. guttatus*; MTRU, *M. truncatula*; PITA, *P. taeda*; PTRI, *P. trichocarpa*; SLYC, *S. lycopersicum*; UGIB, *U. gibba*; VVIN, *V. vinifera*. Details of gene families in column two are inferred from the gene family membership/function of *A. thaliana* genes (according to The Arabidopsis Information Resource; http://www.arabidopsis.org) belonging to these clusters. It should be noted that OrthoMCL clusters are not necessarily equivalent to gene families as a single gene family may be split over multiple clusters and multiple gene families may be grouped into a single cluster.