Europe PMC Funders Group
Author Manuscript
Nat Ecol Evol. Author manuscript; available in PMC 2020 May 18.

Published in final edited form as:
Nat Ecol Evol. 2019 December ; 3(12): 1686–1696. doi:10.1038/s41559-019-1036-6.

Genomic basis of European ash tree resistance to ash dieback fungus

Jonathan J. Stocks[1,2], **Carey L. Metheringham**[1,2], **William J. Plumb**[1,2,3], **Steve J. Lee**[4], **Laura J. Kelly**[1,2], **Richard A. Nichols**[1], **Richard J. A. Buggs**[1,2,*]

[1]School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS, UK [2]Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AE, UK [3]Forestry Development Department, Teagasc, Dublin, Republic of Ireland [4]Forest Research, Northern Research Station, Roslin Midlothian, EH25 9SY, UK

Summary

Populations of European ash trees (*Fraxinus excelsior*) are being devastated by the invasive alien fungus *Hymenoscyphus fraxineus*, which causes ash dieback (ADB). We sequenced whole genomic DNA from 1250 ash trees in 31 DNA pools, each pool containing trees with the same ADB damage status in a screening trial and from the same seed-source zone. A genome-wide association study (GWAS) identified 3,149 single nucleotide polymorphisms (SNPs) associated with low versus high ADB damage. Sixty-one of the 192 most significant SNPs were in, or close to, genes with putative homologs already known to be involved in pathogen responses in other plant species. We also used the pooled sequence data to train a genomic prediction model, cross-validated using individual whole genome sequence data generated for 75 healthy and 75 damaged trees from a single seed source. Using the top 20% of our genomic estimated breeding values from 200 SNPs, we could predict tree health with over 90% accuracy. We infer that ash dieback resistance in *F. excelsior* is a polygenic trait that should respond well to both natural selection and breeding, which could be accelerated using genomic prediction.

Keywords

Fraxinus excelsior; ash; *Hymenoscyphus fraxineus*; ash dieback; pool-seq; Genome wide association study (GWAS); Genomic Selection (GS)

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Correspondence: r.buggs@kew.org.

Author Contributions

J.J.S. performed the field assessments and sampling, data analysis for all the GWASs, GS for dataset B and wrote the manuscript. R.J.A.B supervised field work, data analysis and interpretation and wrote the manuscript. L.J.K. analysed genetic data. S.J.L designed the field trials. R.A.N designed the statistical approaches. C.L.M developed and performed methods for Genomic Prediction with training on pool-seq data. W.J.P. modelled the proteins. All authors reviewed the manuscript.

Declaration of Interests

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Introduction

Fraxinus excelsior (European ash), is a broad-leaved tree species widespread in Europe, with 953 ecologically associated species in the UK[1], and with high genetic diversity[2]. Its populations are being severely reduced by the invasive alien fungus *Hymenoscyphus fraxineus*, which causes ash dieback[3,4]. Several previous studies have shown that there is a low frequency of heritable resistance to ADB in European ash populations[5]. Estimates of breeding values of mother trees based on observed ADB damage in their progeny have an approximately normal distribution, hinting that resistance is a polygenic trait[6] that would respond well to selection. An associative transcriptomics study on 182 Danish ash trees found expression levels of 20 genes associated with ADB damage scores but no genomic SNPs[2]. In model organisms, crops and farm animals, analysis of genomic information has been widely used to discover candidate genes involved in phenotypic traits, or to identify individuals with desirable breeding values[7–13]. The identification of candidate loci typically makes use of genome-wide association studies (GWAS) whereas genomic prediction (GP) methods can be used to select individuals with high breeding values. These methods have seldom been applied to keystone species in natural ecosystems due to the typically high genetic variability of such species and the high cost of genome-wide genotyping. Previous studies have demonstrated that estimation of allele frequencies by sequencing of pooled DNA samples (pool-seq) can reduce the cost of a GWAS[14], but thus far such data have not been applied to the training of GP models. Here, we applied pool-seq GWAS and pool-seq trained GP models to European ash populations, finding a large number of SNPs associated with ADB damage that allow us to make accurate estimates of breeding values (Extended Data Fig. 1).

Results

Genome-wide association study

For 1250 ash trees we generated average genome coverage of 2.2x per tree, within DNA pools of 30-58 trees (Supplementary Table 1). Each pool contained DNA from trees from one of thirteen geographical seed source zones, and from trees that were either healthy or highly damaged by ADB in a mass screening trial[15] (Supplementary Table 2). On average 98.3% of reads per pool mapped to the ash reference genome assembly[2] (Supplementary Table 1). After filtering read alignments for quality, coverage, indels and repeats, we calculated allele frequencies at 9,347,243 SNP loci. A correspondence analysis (CA), on the major allele frequencies for all 31 pools showed a distribution reflecting the geographic origin of the seed sources (Fig. 1), in which axis 1 (summarising 10% of variation) reflected latitude and axis 2 (summarising 9% of variation) reflected longitude. Allele frequency measures were highly correlated in technical and biological replicates (Extended Data Fig. 2). We carried out a GWAS of allele frequencies in healthy versus ADB-damaged pools paired by seed source zone using a Cochran-Mantel-Haenszel (CMH) test. We excluded 15,739 SNPs (0.17% of the 9,347,243 SNP loci) that were found in contaminant contigs comprising 0.50% of the reference genome (Extended Data Fig. 3). We found 3,149 SNP loci significantly associated with ash dieback damage level with a local FDR cut-off at 1x

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

10^{-4} (Supplementary Table 3, Extended Data Fig. 4). Imposing a more stringent cut-off of 1 x 10^{-13}, we found 192 significant SNP loci (Fig. 2).

Seven genes contained missense variants caused by ten of these 192 SNPs (Table 1, Fig. 3, Supplementary Table 5). We were able to model the proteins encoded by four of these genes (Extended Data Fig. 5). Similarity searches on these seven genes suggested that four of them are already known to be involved in stress or pathogen responses in other plant species. Gene FRAEX38873_v2_000003260, is putatively homologous to an *Arabidopsis* BED finger-NBS-LRR-type Resistance (R) gene (At5g63020)[16] and is affected by a leucine/tryptophan variant close to the protein's nucleotide binding site (Extended Data Fig. 5a) with the tryptophan being rarer overall, but at a higher frequency in the healthy than the damaged trees (Supplementary Table 5). This R gene is located (see Fig. 3b) on Contig 10122 less than 5Kb from gene FRAEX38873_v2_000003270, which is putatively homologous to a Constitutive expresser of Pathogenesis-Related genes-5 (CPR5)-like protein and affected by an isoleucine/serine variant, a 5' UTR start codon variant and 16 non-coding variants. This CPR5-like gene is likely to regulate disease responses via salicylic acid signalling[17]. Gene FRAEX38873_v2_000164520 is a putative F-box/kelch-repeat protein SKIP6 homolog, which encodes a subunit of the Skp, Cullin, F-box containing (SCF) complex, catalysing ubiquitination of proteins prior to their degradation[18]. One of our candidate SNPs encodes an arginine/glutamine substitution in this gene, with the arginine being rarer overall, but at a higher frequency in the healthy than the damaged trees. The substitution is located close to the gene's F-box motif (Extended Data Fig. 5b) and is likely to affect binding within the SCF complex due to the charge difference between the two amino acids. In pine trees, F-Box-SKP6 proteins have been linked to fungal resistance[19]. Gene FRAEX38873_v2_000305440, may also be involved in ubiquitination: although the CDS hit an uncharacterised gene in olive (Table 1), the mRNA hit an E3 ubiquitin-protein ligase. This gene contains a glycine to aspartic acid substitution.

The other three genes with missense mutations have putative homologs with functions that have not been previously linked directly to disease resistance. Gene FRAEX38873_v2_000116110 is a 60S ribosomal protein L4-1 (RPL4-1) homolog, with four missense and nine synonymous variants associated with ADB damage level. The amino acid positions affected are in disordered regions in close proximity to one another (Extended Data Fig. 5d). Changes in this gene may affect the efficiency of mRNA translation[20]. Gene FRAEX38873_v2_000346660 is a Heat Intolerant 4 like protein with a phenylalanine to leucine variant. Gene FRAEX38873_v2_000180950 is a homolog of Damaged DNA-Binding 2 (DBB2), which has a role in DNA repair[21] and contains a proline/leucine substitution within its WD40 protein binding domain (Extended Data Fig. 5c). This gene is found on Contig 332 between two G-type lectin S-receptor-like serine/threonine-protein kinase LECRK3 genes (FRAEX38873_v2_000180940 and FRAEX38873_v2_000180960) whose putative homologs are involved in brown planthopper resistance in rice[22].

A further 24 genes contain significant ($p < 1 \times 10^{-13}$) SNPs encoding variants that are transcribed but not translated (Table 1) and may therefore affect expression of these genes. Of these, four match genes that have been previously identified as involved in disease resistance in other species. Gene FRAEX38873_v2_000234590 encodes a WPP domain-

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

interacting protein 1-like, and WPP domains have been linked to viral resistance in potato[23]. Gene FRAEX38873_v2_000305460 encodes a PHR1-LIKE 3-like protein which may play a role in immunity[24] via the salicylic acid and jasmonic acid pathways[25]. Gene FRAEX38873_v2_000013250 encodes a Membrane Attack Complex and Perforin (MACPF) domain-containing Constitutively Activated cell Death (CAD) 1-like gene, which controls the hypersensitive response via salicylic acid dependent defence[26]. FRAEX38873_v2_000211580 is a Squalene monooxygenase-like gene involved in the synthesis of phytosterols [27], which have a role in plant immunity[28].

Other genes involved in regulation were found to have significant ($p < 1 \times 10^{-13}$) non-translated variants. FRAEX38873_v2_000266510 is a zinc finger CCCH domain-containing protein 11-like that is likely to be involved in regulation, perhaps of resistance mechanisms[29]. FRAEX38873_v2_000047060 is a short-chain dehydrogenase TIC 32, chloroplastic-like gene that is involved in the regulation of protein import[30]. FRAEX38873_v2_000074310 is putatively homologous to a squamosa promoter-binding (SBP)-like protein 8 that controls stress responses in *Arabidopsis*[31]. Two genes with non-coding variants seem to affect phenology: gene FRAEX38873_v2_000145630 encodes a Vernalisation Insensitive 3 (VIN3) like protein 1[32] and gene FRAEX38873_v2_000168770 encodes a Late Flowering-like protein. A further two intron variants were located on another putative DNA repair gene (in addition to FRAEX38873_v2_000180950, which had a missense variant); gene FRAEX38873_v2_000308800 encoding a probable DNA helicase MiniChromosome Maintenance (MCM) 8 protein.

Six genes with putative roles in disease resistance have significant ($p < 1 \times 10^{-13}$) SNPs within 5Kb up- or down-stream of them and are the closest known genes to those SNPs (Table 1). FRAEX38873_v2_000296810 matches an ankyrin repeat-containing protein NPR4-like gene; in *Arabidopsis* the *NPR4* gene is involved in defence against fungal pathogens and in mediation of the salicylic acid and jasmonic acid/ethylene-activated signalling pathways[33]. FRAEX38873_v2_000190500 is a putative ethylene-responsive transcription factor ERF098-like gene which may be involved in regulation of disease resistance pathways[34]. Gene FRAEX38873_v2_000342260 is a palmitoyltransferase or protein S-acyltransferases (PATs) 8-like gene[35], which is likely to have a role in protein trafficking and signalling; in *Arabidopsis*, some PATs regulate senescence via the salicylic acid pathway[36]. FRAEX38873_v2_000025560 encodes a probable xyloglucan endotransglucosylase/hydrolase protein 27 which may play a role in extracellular defence against pathogens[37,38]. FRAEX38873_v2_0000258470 encodes an F-box/FBD/LRR-repeat protein likely to be involved in ubiquitination (see above). FRAEX38873_v2_0000340820 is a putative dehydration-responsive element-binding protein 2C-like (DREB2C) gene which has a role in osmotic-stress signal transduction pathways[39].

For 49 of the 192 most significant GWAS SNPs ($p < 1 \times 10^{-13}$), their closest gene was between 5Kb and 100Kb distant; these were identified by SNPeff as “intergenic SNPs” (Table S4). These included some with previous evidence of disease resistance functions. Gene FRAEX38873_v2_000086110 is a Leucine-rich repeat receptor-like serine/threonine-protein kinase β-amylase (BAM) 3, which is involved in fungal resistance in *Arabidopsis* [40]. Gene FRAEX38873_v2_000291580 is a bHLH162-like transcription factor whose putative

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Arabidopsis homolog is induced by infection with the downy mildew pathogen *Hyaloperonospora arabidopsidis* [41]. Gene FRAEX38873_v2_000169770 is likely to be involved in vacuolar protein sorting which can play a role in defence responses[42]. A cluster of SNPs on contig1355 are located at approximately 13-kb from gene FRAEX38873_v2_000037990, a small ubiquitin-like modifier (SUMO) conjugating enzyme UBC9-like gene. Inhibition of SUMO conjugation in *Arabidopsis* causes increased susceptibility to fungal pathogens[43]. Gene FRAEX38873_v2_000282910 is a nitrate regulatory gene 2 (NRG2) which could mediate nitrate signalling or mobilisation[44]. Gene FRAEX38873_v2_000340830 is a trichome birefringence-like (TBL) 33 gene; mutants of TBL genes in rice plants confer reduced resistance to rice blight disease[45].

Genomic prediction

We individually sequenced from the same trials 150 trees that had not been included in the DNA pools. These 150 trees were 75 healthy and 75 unhealthy trees from seed-source NSZ 204. For them we generated a total of 2.9Tbp data in 19.5 billion reads (Dataset B). Each individual tree was sequenced to 22X genome coverage on average. Quality metrics and GC content were very similar to Dataset A (Supplementary Table 1). On average the percentage of reads mapped to the reference genome assembly per sample was 98.4% and 32,443,401 SNPs were found with read depth > 9 and mapping quality > 15.

To evaluate the genomic estimated breeding values (GEBV) of ADB damage, we used the pool-seq data as a training population and the 150 NSZ 204 individuals as a test population. We obtained highest accuracy (correlation of observed scores and GEBV, $r = 0.35$; frequency of correct allocations, $f = 0.67$) using the top 10,000 SNPs by p-value from the GWAS, of which 9,620 SNPs had been successfully called in the test population (Fig. 4). Smaller and larger SNP-dataset sizes performed less well. With a view to using a subset of these SNP for prediction, we reran the analysis using a subsets of SNPs with the largest (absolute) estimated effect sizes and observed a small increase in correlation (Fig. 4), finding the best result with 25% of the dataset of 10,000 SNPs ($r = 0.37$; $f = 0.67$). Estimated effect sizes for all SNPs with models trained on 100 to 50,000 SNPs are shown in Supplementary Table 7c-j.

Using the GWAS p-values as the criterion for selecting candidate SNPs for GP was far more effective than using a random selection from the genome, as judged by r and f scores (Fig. 4). Despite this effect, there was not a strong association between the GWAS p-values and the effect size estimated by the genomic prediction: only 66 of the 2500 SNPs with the largest effect size were in the top 192 SNPs identified by the GWAS.

In a relatively small population with large heritable effects, spurious associations between some SNP alleles and a trait can arise. A sufficiently large number of randomly chosen SNPs will convey all the information on the relatedness of the individuals which, in turn, can be used to predict a trait simply because related individuals have similar trait values. To evaluate this effect, the 150 NSZ 204 individuals were used for GP as both a training dataset and a test dataset. The accuracy of the prediction with the top 50,000 GWAS-identified SNPs was no better than a random selection of 50,000 SNPs (Extended Data Fig. 6). Given this, we re-ran GP training on the pool-seq data with the pools from the same seed source of

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

the test population (NSZ 204) excluded in case their inclusion had given spurious associations that contributed to the success of the first GP. This more stringent cross-validation showed a comparable performance to our previous GP trained on the full pool-seq dataset (maximum $r = 0.36$, maximum $f = 0.67$; Extended Data Fig. 7).

For a breeding programme for increased resistance to ash dieback, accurate prediction of the most resistant trees is needed. We therefore examined the accuracy with which our highest GEBVs were assigning trees correctly to the undamaged health category. For the trees with the top 20% and 30% GEBV scores, we obtained predictive accuracies of $f > 0.9$ and $f > 0.8$ respectively, using as few as 200 predictive SNPs (Fig. 5).

Discussion

Many of the top SNP loci that we found associated with ash tree resistance to ash dieback are in, or close to, genes with putative homologs in other species that have been previously shown to detect pathogens, signal their presence, or regulate pathogen responses. Using SNPs identified by the GWAS to train GP on the pool-seq data, we obtained much greater accuracy in predicting the ADB damage score in 150 separate individuals than when we used the same number of randomly selected SNPs. These results demonstrate we can use genotype to predict performance across different seed-sources, and suggest that other genes that have not previously been implicated in plant pathogen resistance may be involved in resistance to ADB. The distribution of effect sizes and the predictivity peak using 2500 SNPs suggests that *F. excelsior* resistance to *H. fraxineus* is a highly polygenic trait and may therefore respond well to artificial and natural selection, allowing the breeding or evolution of durable increased resistance.

None of our 192 most significant GWAS SNPs were in 20 genes previously identified as gene expression markers (GEMs) associated with ADB resistance[2], but this is not unexpected given that the previous study[2] did not find SNPs associated with ADB resistance in these 20 genes either. Although none of our most significant SNPs had one of these GEMs as their closest gene we cannot exclude the possibility that our candidate SNPs may influence expression of these genes. In any case, the GEMs were identified based on a small sample size of 182 trees[2] and may have been specific to the Danish populations they were sampled from.

The levels of accuracy which our GP reached are high, and comparable to those that are used to inform selections in crop[46–50], tree[12,51] and livestock breeding programmes[52,53]. Thus, our results have the potential to increase the speed at which we can successfully breed ash dieback resistant trees. A common short-coming of GP is that predictions are highly population specific[12,54,55], and the success of GP using randomly selected SNPs when training GP models within the individually sequenced trees suggests that population-specific GP can be easily made for ash. However, we made successful predictions in the individually sequenced trees using the pool-seq trained GP even when the pool-seq data for their seed-source was not used in training the model. This suggests we have successfully identified widespread alleles that are involved in ADB resistance in many populations. There may well be further population-specific alleles that our methods have not detected. Thus, we have

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

used pool-seq data to train a trans-populational GP model. The success of this approach in European ash – a genetically variable species – suggests it may be useful in many other ecologically important species as a cost-effective approach to successful genomic prediction of evolving traits.

Methods

Trial design

This study is based on a Forest Research mass screening trial planted in spring 2013, in areas of high natural *Hymenoscyphus fraxineus* inoculum pressure. The trial comprises 48 hectares of trials on 14 sites in southeast England as described in Stocks *et al.* 2017[15]. Briefly, each site was planted in spring 2013 with two-year-old saplings grown from seed sources from up to 15 different native seed zones (NSZ). These were 10 British NSZ (NSZ 106, NSZ 107, NSZ 109, NSZ 201, NSZ 204, NSZ 302, NSZ 303, NSZ 304, NSZ 403, NSZ 405), Germany (DEU), France (FRA), Ireland (CLARE and IRL DON), and a Breeding Seedling Orchard (BSO) planted by Future Trees Trust (FTT) comprised of half-sibling families from “plus” trees across Britain. Each of the sampled sites had four complete replications. Each site was planted at the high density of 5,000 trees/ha (a spacing of 1 x 2 meters).

Phenotyping and sampling

A survey of the two trial sites with the highest levels of ADB infection (Site 16, near Norwich, Norfolk and Site 35 near Tunbridge Wells, Kent) was carried out in 2016 and is reported in Stocks *et al.* 2017[15]. In July/August 2017 we revisited these sites and collected leaf samples from all trees that were healthy at the time of sampling (score 7 on the scale of Pliura *et al.*[56]). For each healthy tree we sampled, we also sampled a tree with considerable ADB damage (scores 4 or 5 on the scale of Pliura *et al.*[56]). The number of healthy trees at these two sites were insufficient for our experimental design, so we also sampled two other severely affected sites, 21 (near Maidstone, Kent) and 23 (near Norwich, Norfolk). In total we examined 38,784 trees and found only 792 (1.96%) healthy trees. These trees are unlikely to have escaped inoculation, as all had direct neighbours that were diseased and the trees were densely planted. Initially a total of 1536 trees were sampled. Of these, after DNA quantity and quality checks, 623 healthy and 627 damaged trees were selected for pooled sequencing with the total number of trees for each seed source and health status described in Table S2. For individual sequencing, we selected 75 healthy and 75 damaged trees, across the four sampled sites, from a seed source that had a large number of healthy trees (NSZ 204).

DNA extraction and sequencing

Leaf samples were transported to the lab using cool boxes. Fresh Genomic DNA was extracted from liquid nitrogen frozen leaf tissue using the DNeasy Plant Mini Kit or the DNeasy 96 Plant Kit (Qiagen) and eluted in 70 μl of Qiagen AE buffer. Quantification of genomic DNA was performed using the Quantus™ Fluorometer on all extractions. DNA purity quality checks were carried out using the Thermo Scientific™ NanoDrop 2000 for nucleic acid 260/280 and 260/230 absorbance ratios. Of the total number of extractions,

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

1400 were selected based on DNA quantity and quality thresholds. A minimum concentration of >20 ng/μl, OD260/280 >1.7 and total amount >1.0 μg of DNA was necessary for the sample to pass. Of the 1400 samples, 1250 were separated for the pooling and sequencing procedures and will be referred to as dataset A. A separate 150 individuals from NSZ 204, that were not included in the pools, were selected for individual genotyping and will be referred to as dataset B.

For the pooling procedure equal amounts of DNA from each sample were pooled together based on their initial DNA concentrations, adjusting the total volume of each sample accordingly. Pooling was based on seed source origin and health status with two pools for each seed source, one healthy and the other damaged. A total of 31 pools were created (Supplementary Table 2), one being a technical replicate of the healthy trees from NSZ 204 that was made by independently repeating all quantification, quality and pooling steps on the same 40 trees. NSZ 106 and NSZ 107 had 4 pools each as the samples were divided to maintain an average of 42 trees per pool. These therefore provide biological replicates. Studies have shown that pools sizes as small as 12 have provided robust and reliable population allele frequency estimates[14,57].

TruSeq DNA PCR-Free (Illumina) sequencing libraries were prepared, using 350 base pair inserts. All sequencing was carried out using HiSeq X at Macrogen (South Korea) with 150 paired end reads with the goal of achieving a whole genome coverage (based on the estimated genome size of the *F. excelsior* reference individual[2] of 80x per pool (2x coverage per individual) for dataset A and 20x for dataset B.

Mapping to reference and filtering

Trimmomatic v0.38[58] was used for read trimming and adapter removal. Leading and trailing low quality or N bases below a quality of 3 were removed. Reads were scanned with a 4-base wide sliding window, cutting when the average quality per base dropped below 15 and excluding reads below 36 bases long[58]. Reads were then aligned to the reference genome for *Fraxinus excelsior*, assembly version BATG0.5[2], using the Burrows-Wheeler Alignment Tool (BWA MEM)[59], v. 0.7.17 with default settings. The mapped reads were filtered for a mapping quality of 20 with SAMtools v1.9[60]. On average the percentage of reads mapped to the reference was 98.3% for dataset A and 98.4% for dataset B. For both datasets Sequence Alignment Map (SAM) and binary version (BAM) files were created using SAMtools. Indels were detected and removed using PoPoolation2[61] scripts (identify-indel-regions.pl and filter-sync-by-gtf.pl) that include five flanking nucleotides on both sides of an indel. The position of repeats in the reference genome was annotated previously[3] using RepeatMasker v. 4.0.5 (with option -nolow) and that information used to remove repeats from these data using the same removal script provided by PoPoolation2.

Genetic structure of seed sources

Major allele frequency information was extracted from dataset A for each of the 31 populations using a modified output of the allele frequency differences script (snp-frequency-diff.pl) from the PoPoolation2 package. This table of major allele frequencies was imported and converted to a genpop object and subsequently analysed using the R package

adegenet[62] by performing a Correspondence Analysis in order to seek a typology of populations. Correlation between populations was calculated and plotted, for the major allele frequencies from dataset A, using the corrplot R package[63].

Genome wide association study

Dataset A was analyzed using the software package PoPoolation2[61] in a genome wide association study (Extended Data Fig. 1). An mpileup input was generated using SAMtools followed by the creation of a file that had all the variants synchronized across the pools, requiring a base quality of at least 20. The Cochran-Mantel-Haenszel (CMH) test[64] was used to identify significant and consistent allele frequency differences between damaged and healthy trees, with each seed source pair used as an independent measurement. The technical replicate of NSZ 204 was not used, and the biological replicates of NSZ 106 and NSZ 107 were treated as independent measurements. Thus, a 2x2 data table was created for each SNP locus in each pair of pools. The counts of each allele for each phenotype were treated as the dependent variables. The parameters set for PoPoolation2 were: min count 15 (minimum allele count to be included), min coverage 40, max coverage 3000. The "-- population" option was used to define the pair-wise comparisons between the pools from each seed source. False discovery rate control was performed using the R package q-value[65].

Contaminant sequences were detected using Blobtools v1.1[66]. This used three input files: the reference assembly fasta file (BATG0.5), a coverage file and a hits file. The coverage file was a mapping to BATG0.5 of paired 100bp Illumina reads with insert sizes of 200bp, 300bp and 500b that were used in the original assembly of BATG0.5[2] using Bowtie 2 v.2.3.0 with the "very-sensitive" preset and setting "maxins" to 1000. The mapping was converted to BAM format and sorted using the "view" and "sort" functions in SAMtools v.1.4.1. The hits file was a BLAST+ output for all contigs in the *F. excelsior* reference assembly with the top score results in the outfmt 6 format including fields "qseqid sseqid staxids bitscore". Blobtools function "create" was used to assign a taxonomy under a given taxonomic rule to each sequence in the assembly. NCBI nodes and names files were provided to infer the taxonomy at each rank. Of the 89,514 scaffolds and contigs in the BATG0.5 genome assembly, 2,408 short contigs appeared to be contaminant as they showed a phylum taxonomic rank different to Streptophyta (Extended Data Fig. 3, Supplementary Table 7a).

Putative functions for genes containing, or near, the pool-seq GWAS top 192 SNPs were assigned by obtaining the CDSs from the Ash Genome website[2] and using the command line NCBI Basic Local Alignment Search Tool (BLAST+) optimized for the megablast algorithm to search the GenBank Nucleotide database. The top result for every BLAST search was extracted and their predicted gene functions were used to functionally annotate the ash genes. Any search that yielded no matches when using megablast was then repeated using the blastn algorithm and ultimately cDNA sequences if the latter was also uninformative. Potential functional impacts for each of the top 192 GWAS SNP loci were determined using SNPeff (v. 4.3T)[67]. A custom genome database was built from the *F. excelsior* reference assembly using the SnpEff command "build" with option "-gtf22"; a gtf file containing the annotation for all genes, as well as fasta files containing the genome assembly, CDS and

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

protein sequences, were used as input. Annotation of the impact of the 192 SNPs was performed by running SnpEff on all *F. excelsior* genes with default parameter settings.

Protein modelling

Proteins containing SNPs identified by SnpEff as coding for amino acid substitutions were modelled. Protein coding sequences were taken from the predicted proteome of the BATG 0.5 reference genome[2] and modelled both with the amino acid(s) associated with ADB damage in our GWAS, and with the amino acid(s) associated with healthy trees. Models were predicted using three *in silico* methods: RaptorX-Binding (http://raptorx.uchicago.edu/BindingSite/), SWISS-MODEL[68] and Phyre2[69] (for the full list of wwPDB proteins selected and used as templates by Phyre see Supplementary Table 6). These models were compared by using the align function in PyMOL v.2.0[70], and only those with congruent models were taken forward, based on their Phyre2 and RaptorX-Binding models. Potential binding sites and candidate ligands were analysed using RaptorX-Binding and literature searches. SDF files for candidate ligands were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov) and converted to 3d pdb files using Online SMILES Translator and Structure File Generator (https://cactus.nci.nih.gov/translate/). Docking with our protein models was analysed using Autodock Vina v.1.1.2[71] with the GUI PyRx v.0.8[72]. Following docking, ligand binding site coordinates were exported as SDF files from Pyrex and loaded into PyMOL with the corresponding protein model file for the “healthy” and “damaged” protein models. Binding sites were then annotated and the variable residues were labelled. Possible RNA and DNA binding sites were predicted using DRONA (http://crdd.osdd.net/raghava/drona/links.php). The presence of signal peptides were detected using SignalP 4.1 server and Phobius server (http://phobius.sbc.su.se/index.html); both were run with default parameters and for Phobius the “normal prediction” method was used. The presence of a signal peptide was confirmed only if it was predicted by both methods. Motif search (https://www.genome.jp/tools/motif/) and ScanProsite (https://prosite.expasy.org/scanprosite/) were used to predict protein domains and their locations for our candidate genes.

Genomic Prediction

We trained a GP model based on the pool-seq data (Dataset A) excluding contaminant SNPs. Subsets of 100, 200, 500, 1000, 5000, 10000, 25000 and 50000 SNPs with the most significant GWAS results were selected from Dataset A and used as a training set. Results were compared with SNP sets of the same size drawn at random from the genome. We constructed a pipeline available at https://github.research.its.qmul.ac.uk/btx330/gppool. The vector of ADB damage scores for each pool, y, was predicted by the rrBLUP model as: $y = \mathbf{X}\beta + \varepsilon$, where β is a vector of allelic effects (treated as normally distributed random effects), and the residual variance is Var[ε]. The genetic data are encoded in the design matrix **X** which has a row for each pool and a column for each SNP allele. The entry for pool p and locus l is $X[p,l] = f_{pl} - \mu_s$, where f_{pl} is the frequency of the focal allele and μ_s is its mean frequency across the pools from the same seed-source as p.

The Reduced Maximum Likelihood solution to the model was obtained using the *mixed.solve* function in rrBLUP v4.6[73] to give estimated effect sizes (EES) for the minor and major alleles at each SNP under consideration. Subsets of the 10 – 50,000 SNPs with

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

the greatest EES were used to predict GEBV for each of the 150 individuals from NSZ 204. For these individuals (dataset B) variant calling was performed using BCFtools with the raw set of called SNPs filtered using VCFtools (vcfutils) - set at minimum read depth of 10 and minimum mapping quality 15. Filtering of loci was carried out using thresholds of >95% call rate and >5% MAF. Samples were filtered based on a >95% call rate and <1% inbreeding coefficient. SNPs were also filtered if they deviated significantly from Hardy-Weinberg equilibrium. GEBV was calculated as the sum of the EES and the relative frequency of each focal allele. Predictions were repeated with seed-source NSZ 204 excluded from the training dataset to avoid spurious correlations due to population stratification.

Test trees were assigned to high and low susceptibility groups based on their GEBV and the accuracy of the assignment was tested using the formula: f = correct assignments/total assignments, with correct assignments defined as those that corresponded to the observed phenotypes. Correlation of GEBV and phenotypic classification, r, was calculated using the Pearson correlation coefficient.

We also carried out genomic prediction based solely on the 150 individuals in Dataset B. A ratio of 60/40 was used for training and testing populations and missing markers were imputed using the function R package A.mat[74] with default settings. SNPs were selected from the GWAS output ordered by p-value. A total of 100, 500, 1000, 5000, 10000, 50000, 100000, 250000, 500000, 1000000 and 5000000 SNPs were selected from each filtered set and used for training and testing of the GP model. The same number of SNPs were selected at random (using R) from the fully filtered dataset and also used for training and testing the GP model. We used the *mixed.solve* function in rrBLUP v4.6 and Genomic Selection in R course scripts available at http://pbgworks.org. A total of 500 iterations were run of the rrBLUP. For the randomly selected SNPs, the 500 iterations were repeated ten times.

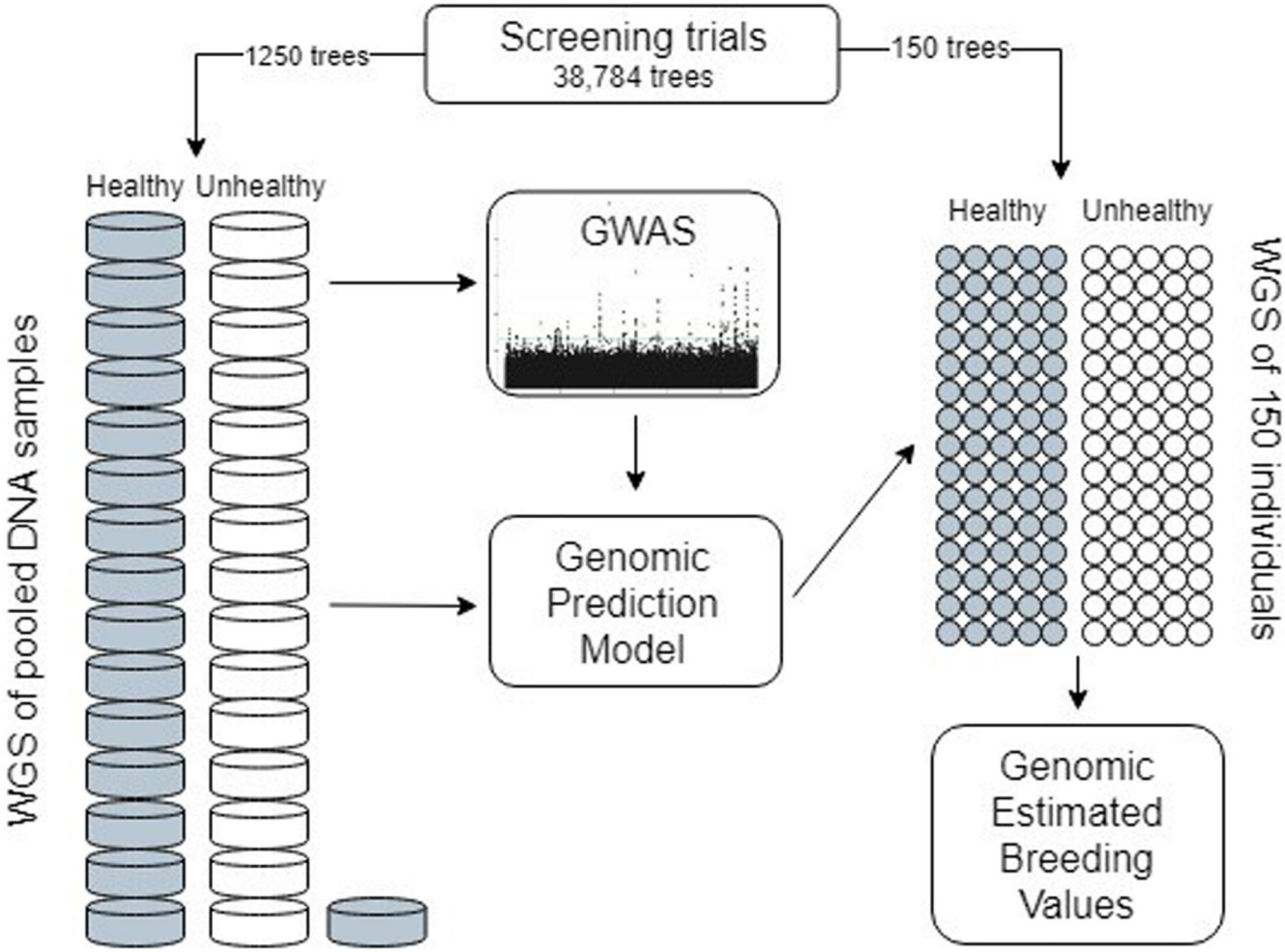
Data, materials and software availability

All trimmed reads are available at the European Nucleotide Archive with primary accession number: PRJEB31096. A guide to these is given in Supplementary Table 7b. The reference *F. excelsior* genome is available for download at www.ashgenome.org and is Assembly GCA_900149125.1 at the European Nucleotide Archive. Biological Materials from the Forest Research Mass Screening trials are available through negotiation of a Materials Transfer Agreement with Forest Research, Northern Research Station, Roslin, Midlothian EH25 9SY. The gppool pipeline developed as part of the project to run GP trained on pool-seq data can be found at https://github.research.its.qmul.ac.uk/btx330/gppool. All software used (Trimmomatic v0.38, BWA MEM v0.7.17, SAMtools v1.9, BCFtools v1.8, VCFtools v 0.1.15, PoPoolation2, R v3.5.3, Repeatmasker v. 4.0.5, Bowtie v. 2.3.0, Blobtools v. 1.1, SNPeff v. 4.3T, Haploview, rrBLUP v4.6, NCBI BLAST, RaptorX-Binding, SWISS-MODEL Phyre2, SMILES, Autodock Vina v.1.1.2, PyRx v.0.8, PyMOL v.2.0, DRONA, SignalP 4.1 server, Phobius server, and NetPhos 3.1 Server) are commercially or freely available.

Extended Data

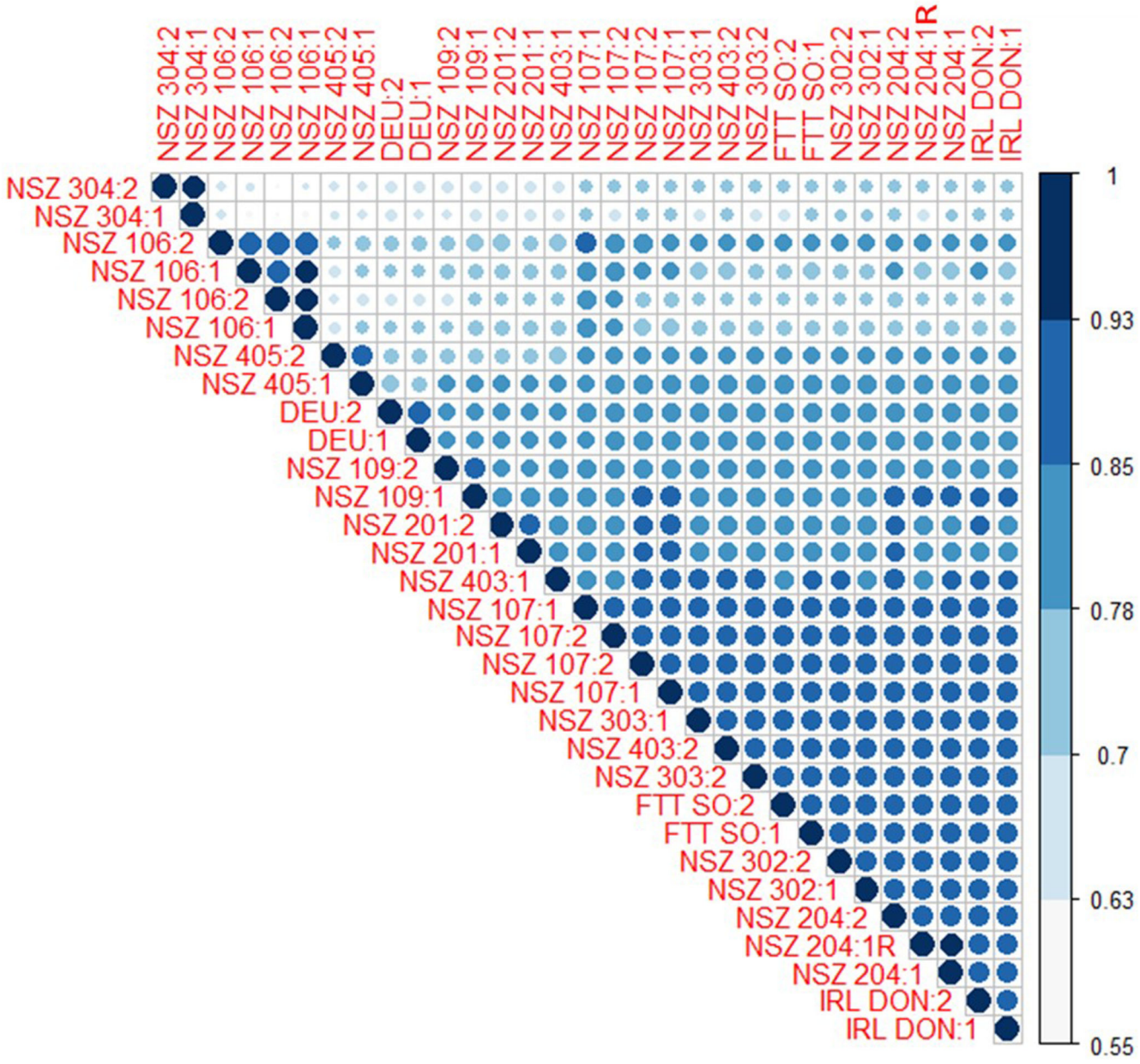
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



Extended Data Fig. 1. Schematic overview of the study design.

Showing sampling and pooling strategies and dependencies of analyses for genome-wide association study and genomic prediction.



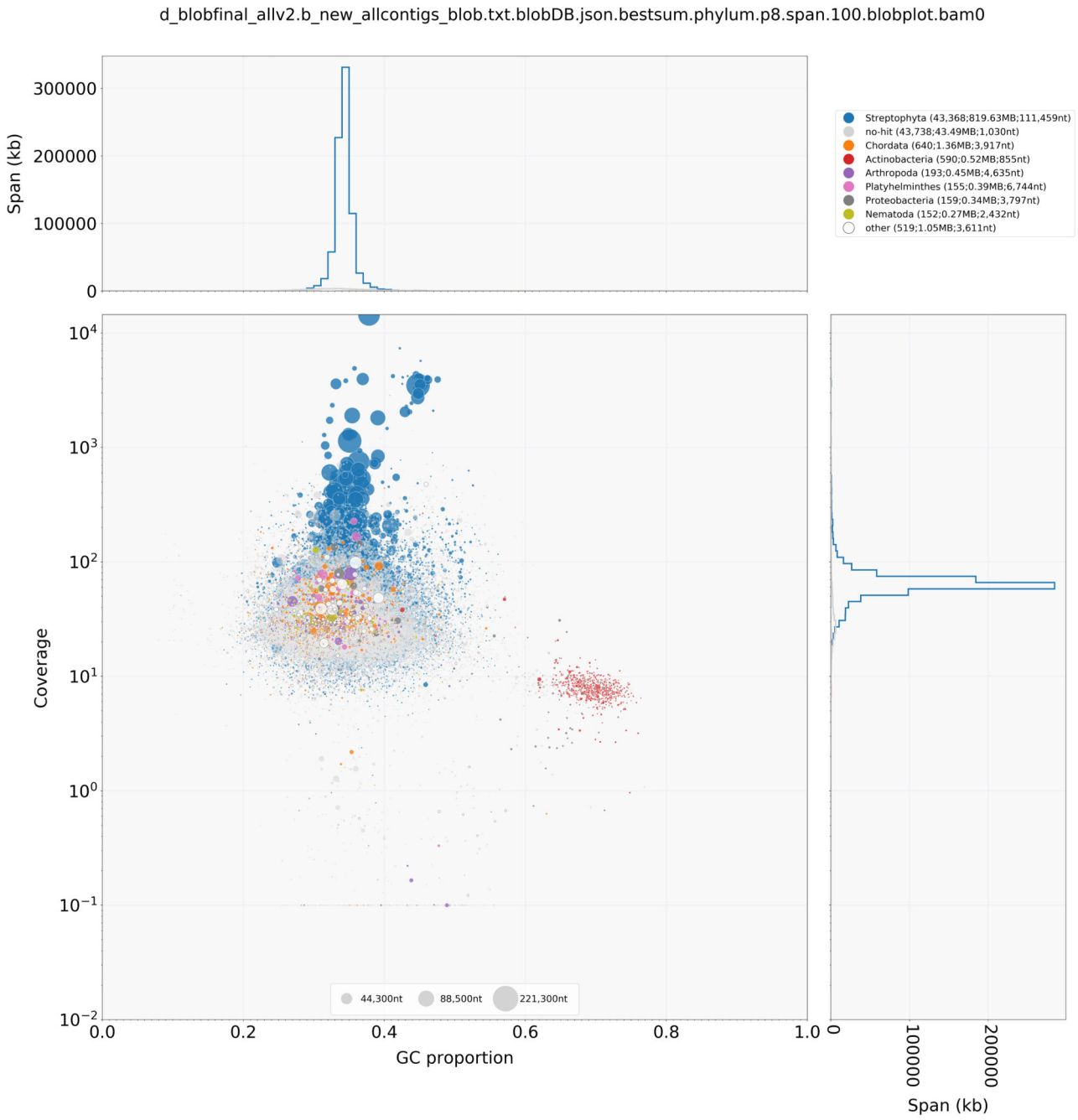
Extended Data Fig. 2. Circle plot of major allele frequency correlation values between all 31 pools in the Pool-seq dataset.

Numbers after seed source code correspond to health status (1 - healthy or 2 - damaged by ADB). Pool NSZ204:1 (with low ADB damage) was technically replicated (NSZ204:1R) using the same set of trees. Both pools from NSZ106 and NSZ107 were biologically replicated for both high and low damage pools, using different sets of trees. High correlation for both technical (NSZ204:1R) and biological replicates (NSZ 106 & 107) can be seen.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

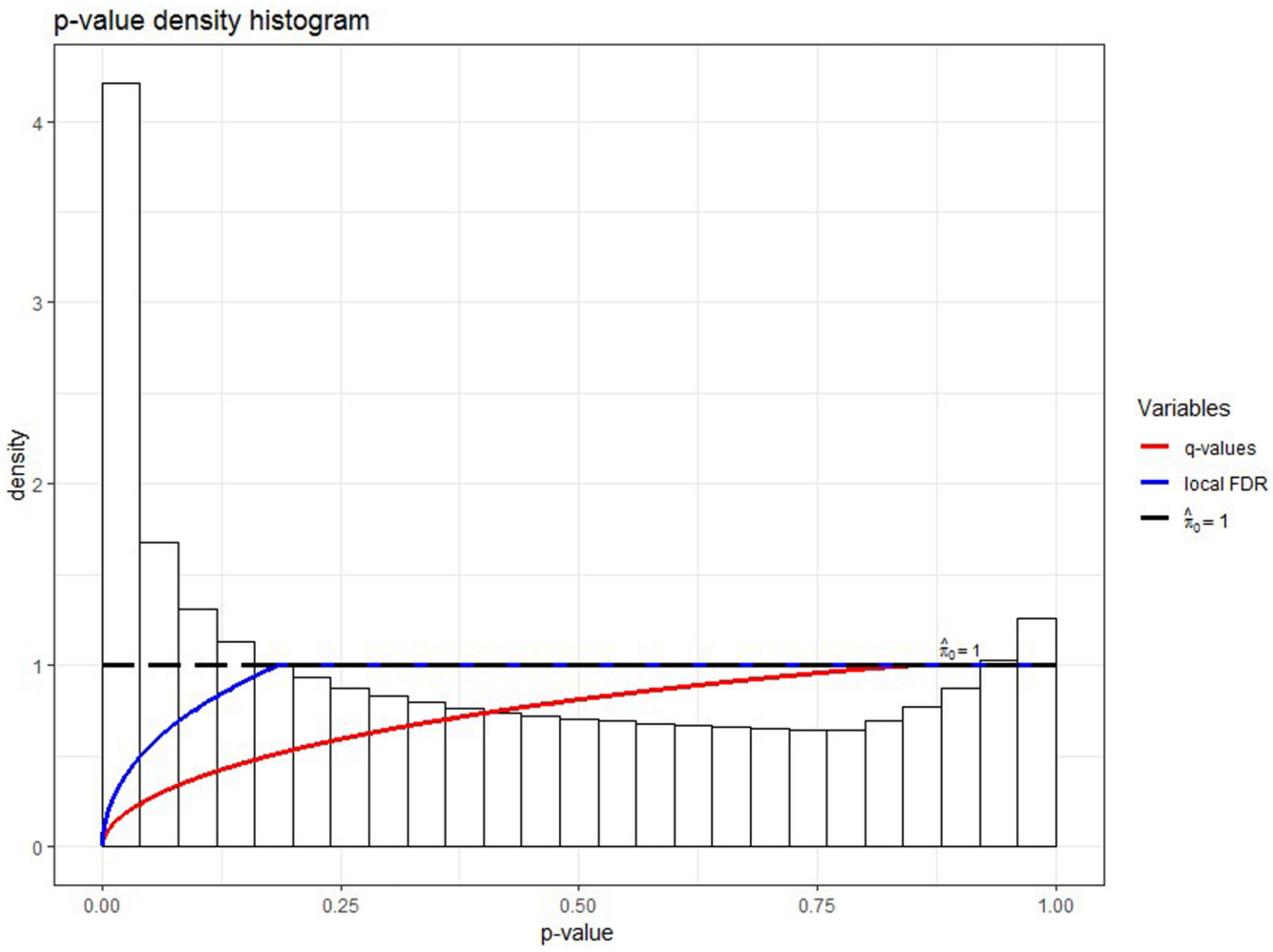


Extended Data Fig. 3. Detection of contamination in the F. excelsior reference genome (BATG0.5).

Blobtools plot for the showing taxonomic affiliation at the phylum rank level, distributed according to GC content and base coverage. Contigs that were not classified as streptophyta corresponded to 0.5% of the genome assembly and 0.24% of all mapped reads.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

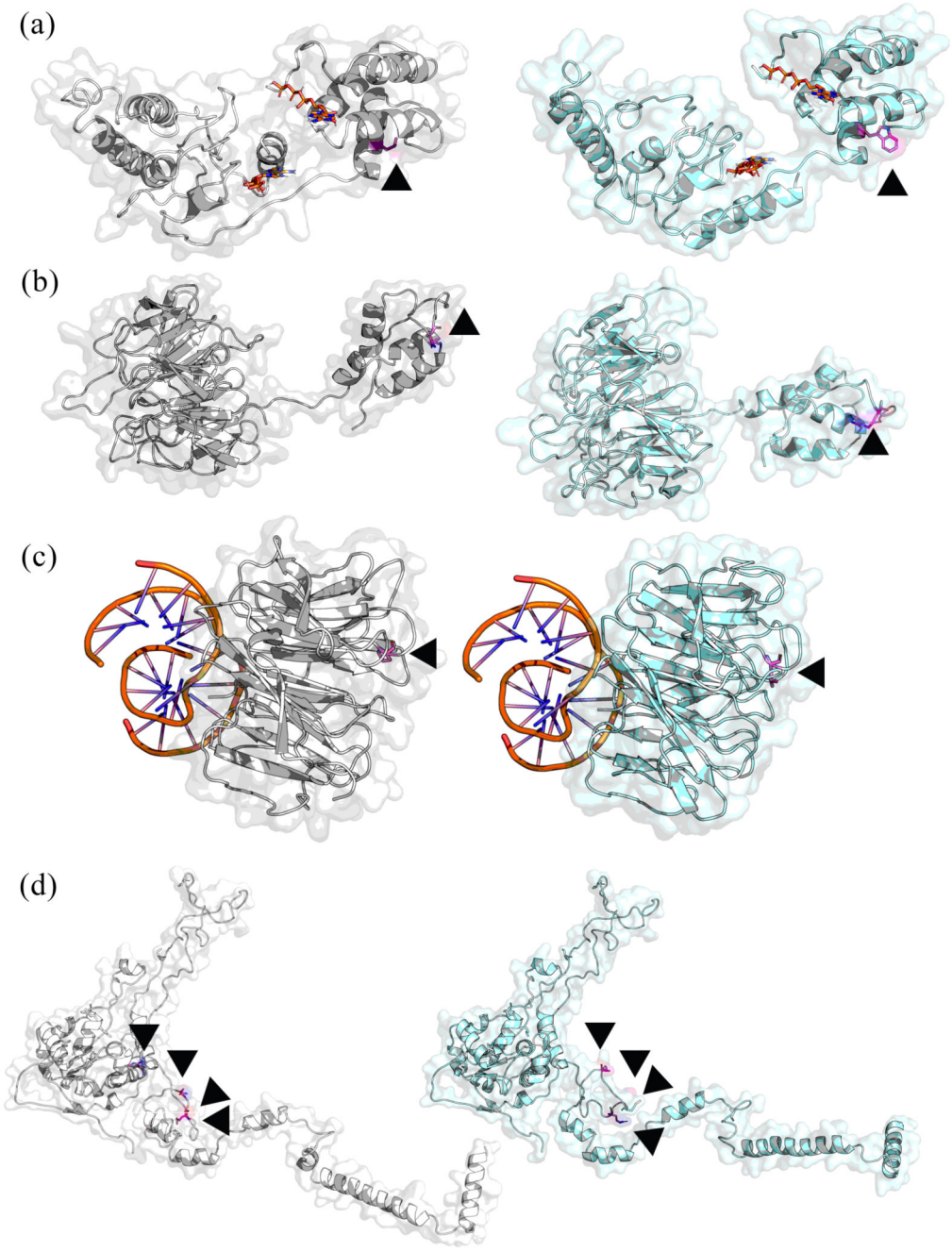


Extended Data Fig. 4. Pool-seq GWAS p-value density histogram with line plots of the q-values and local False Discovery Rate (FDR) values versus p-values.

The π0 estimate is also displayed.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



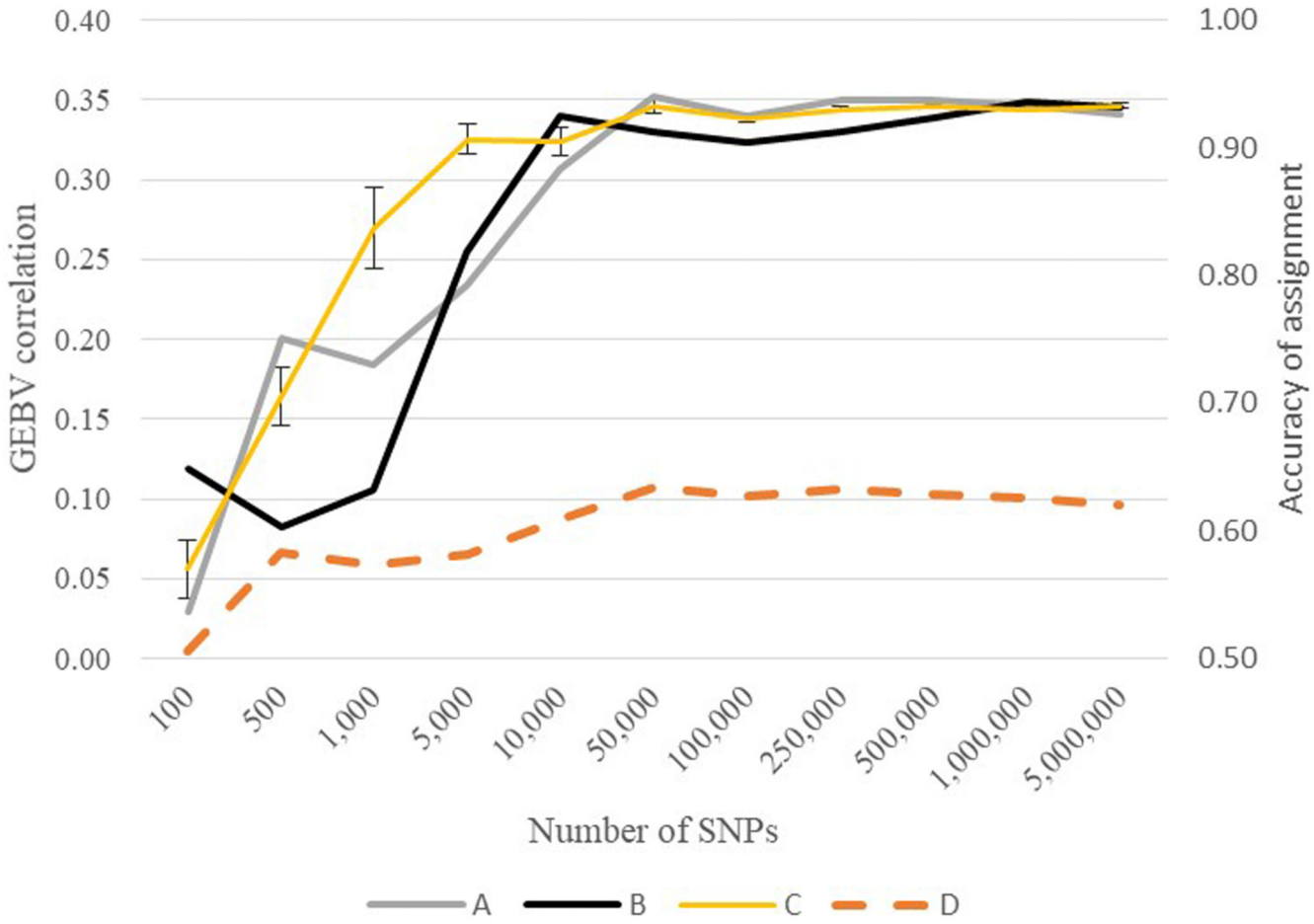
Extended Data Fig. 5. Predicted protein structures for genes containing amino acid changes associated with tree health status under ADB pressure.

The protein structures to the left were more common in damaged trees, and those to the right were more common in healthy trees. Variant amino acids are coloured in magenta and indicated with a black arrowhead. (a) Gene FRAEX38873_v2_000003260, a BED finger-NBS-LRR resistance protein, where position 157 is a leucine (left) versus tryptophan (right) variant. Two ATP molecules are shown in orange to indicate the location of nucleotide binding sites. (b) Gene FRAEX38873_v2_000164520, a F-box/kelch-repeat, where position 13 is a glutamine (left) versus arginine (right) variant. (c) FRAEX38873_v2_000180950, a

Protein DAMAGED DNA-BINDING, where position 99 is a proline (left) versus leucine (right) variant. DNA molecules are shown in orange docked at the proteins' DNA binding sites. (d) Gene FRAEX38873_v2_000116110, a 60S ribosomal protein L4-1, where position 251 is an arginine (left) versus glycine (right) variant, position 285 is a methionine (left) versus arginine (right) variant, position 287 is an asparagine (left) versus lysine (right) variant and position 297 is a threonine (left) versus alanine (right) variant.

Europe PMC Funders Author Manuscripts

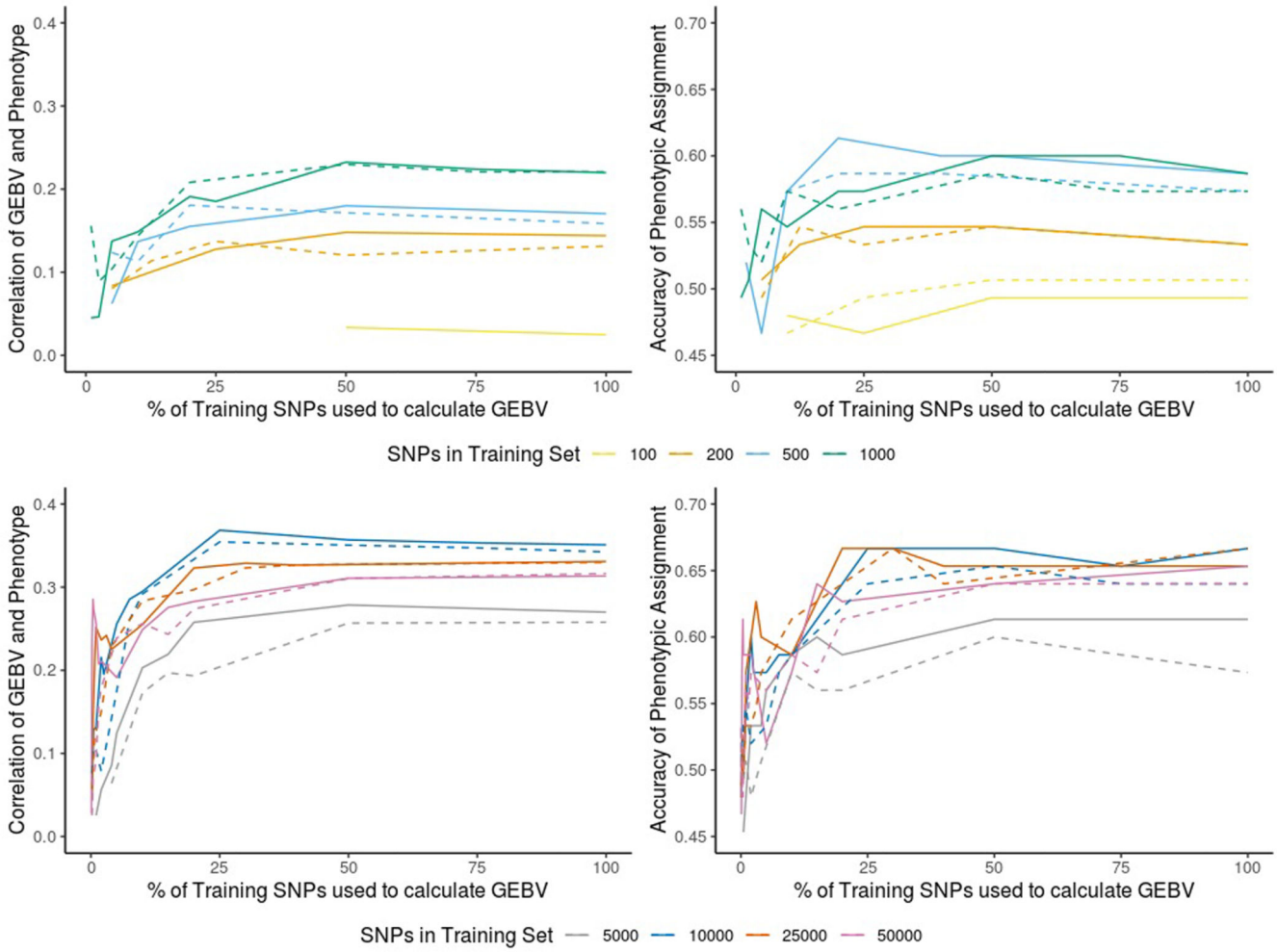
Europe PMC Funders Author Manuscripts



Extended Data Fig. 6. Genomic prediction results using the 150 individually genotyped samples as both training and testing set, showing little difference in accuracy between GWAS SNPs and random SNPs.

(A) GWAS candidate SNPs with all data filters applied (mapping quality, indel and repeat removal); (B) GWAS candidate SNPs only filtering by mapping quality and indel removal; (C) random selection of SNPs using all data filters (mean and standard error shown for N=10 runs, each of 500 iterations); (D) GP allocation accuracy calculated using data with all filters applied. The scale on the left hand vertical axis is for correlation, and the scale on the right hand vertical axis is for accuracy. 100 to 5 million SNPs used to train and test the rrBLUP model.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts



Extended Data Fig. 7. Genomic prediction using Pool-seq data for training and 150 NSZ 204 individuals for testing.

Dashed lines show results excluding Pool-seq data from NSZ 204 (the test seed source) from the training dataset, whereas solid lines show results with NSZ 204 included. The left column shows correlation of observed phenotype and GEBV and the right column shows accuracy of phenotypic assignment from GEBV.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported by Forest Research (FR), Queen Mary University of London (QMUL) and the Royal Botanic Gardens Kew. J.J.S. was funded by a Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) studentship 202790/2014-2 and was part of the Brazilian Scientific Mobility Program – Science without Borders (SwB). S.L. 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 funded jointly by the BBSRC, Defra, Economic and Social Research Council, Forestry Commission, NERC and the Scottish Government. R. J. A. B. and L. J. K were also supported in this work by funding from the Defra Future Proofing Plant Health scheme and the Erica Waltraud Albrecht Endowment Fund. Sequencing was paid for by a direct grant from Defra to RBG Kew. W. J. P. was

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

supported by a Walsh Fellowship from Department of Agriculture, Food and the Marine, Ireland. C. M. was supported by a studentship funded by DEFRA. FR designed and set up the field trials with funding supplied by the Department for Environment, Food and Rural Affairs (DEFRA) contract number TH032 'Rapid screening for Chalara resistance using ash trees currently in commercial nurseries' with additional financial contribution from Department of Agriculture, Food and the Marine, Ireland. The ash trees were all British-grown and sourced from various participating nurseries in England and Scotland. Maelor Forest Nurseries that donated free of charge around half the total number of trees planted.

References

1. Mitchell RJ, et al. Ash dieback in the UK: A review of the ecological and conservation implications and potential management options. Biological Conservation. 2014; doi: 10.1016/j.biocon.2014.04.019
2. Sollars ESA, et al. Genome sequence and genetic diversity of European ash trees. Nature. 2017; doi: 10.1038/nature20786
3. Gross A, Holdenrieder O, Pautasso M, Queloz V, Sieber TN. *Hymenoscyphus pseudoalbidus*, the causal agent of European ash dieback. Mol Plant Pathol. 2014; doi: 10.1111/mpp.12073
4. Pautasso M, Aas G, Queloz V, Holdenrieder O. European ash (*Fraxinus excelsior*) dieback - A conservation biology challenge. Biological Conservation. 2013; doi: 10.1016/j.biocon.2012.08.026
5. Plumb WJ, et al. The viability of a breeding programme for ash in the British Isles in the face of ash dieback. Plants People Planet. 2019
6. Mckinney LV, et al. The ash dieback crisis: Genetic variation in resistance can prove a long-term solution. Plant Pathology. 2014; doi: 10.1111/ppa.12196
7. Endler L, Betancourt AJ, Nolte V, Schlötterer C. Reconciling differences in pool-GWAS between populations: A case study of female abdominal pigmentation in *Drosophila melanogaster*. Genetics. 2016; 202:843–855. [PubMed: 26715669]
8. Fontanesi L, et al. Genome-wide association study for ham weight loss at first salting in Italian Large White pigs: towards the genetic dissection of a key trait for dry-cured ham production. Anim Genet. 2017; doi: 10.1111/age.12491
9. Zhao Y, Mette MF, Gowda M, Longin CFH, Reif JC. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity (Edinb). 2014; 112:638–645. [PubMed: 24518889]
10. Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res (Camb). 2009; doi: 10.1017/S0016672308009981
11. Goddard ME, Hayes BJ, Meuwissen THE. Genomic selection in livestock populations. Genet Res (Camb). 2010; doi: 10.1017/S0016672310000613
12. Müller BSF, et al. Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. BMC Genomics. 2017; 18:1–17. [PubMed: 28049423]
13. Resende JFR, et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics. 2012; doi: 10.1534/genetics.111.137026
14. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014; 15:749–763. [PubMed: 25246196]
15. Stocks JJ, Buggs RJA, Lee SJ. A first assessment of *Fraxinus excelsior* (common ash) susceptibility to *Hymenoscyphus fraxineus* (ash dieback) throughout the British Isles. Sci Rep. 2017; doi: 10.1038/s41598-017-16706-6
16. Bakker EG. A Genome-Wide Survey of R Gene Polymorphisms in *Arabidopsis*. PLANT CELL ONLINE. 2006; doi: 10.1105/tpc.106.042614
17. Meng Z, Ruberti C, Gong Z, Brandizzi F. CPR5 modulates salicylic acid and the unfolded protein response to manage tradeoffs between plant growth and stress responses. Plant J. 2017; doi: 10.1111/tpj.13397
18. Risseeuw EP, et al. Protein interaction analysis of SCF ubiquitin E3 ligase subunits from Arabidopsis. Plant J. 2003; doi: 10.1046/j.1365-313X.2003.01768.x
19. Baker EAG, et al. Comparative Transcriptomics Among Four White Pine Species. G3. 2018; doi: 10.1534/g3.118.200257

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

20. Kakehi JI, et al. Mutations in ribosomal proteins, RPL4 and RACK1, suppress the phenotype of a thermospermine-deficient mutant of arabidopsis thaliana. PLoS One. 2015; doi: 10.1371/journal.pone.0117309
21. Iovine B, Iannella ML, Bevilacqua MA. Damage-specific DNA binding protein 1 (DDB1): A protein with a wide range of functions. International Journal of Biochemistry and Cell Biology. 2011; doi: 10.1016/j.biocel.2011.09.001
22. Liu Y, et al. A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. Nature Biotechnology. 2015; doi: 10.1038/nbt.3069
23. Hao W, Collier SM, Moffett P, Chai J. Structural basis for the interaction between the potato virus X resistance protein (Rx) and its cofactor ran GTPase-activating protein 2 (RanGAP2). J Biol Chem. 2013; doi: 10.1074/jbc.M113.517417
24. Wang S, et al. A noncanonical role for the CKI-RB-E2F cell-cycle signaling pathway in plant effector-triggered immunity. Cell Host Microbe. 2014; doi: 10.1016/j.chom.2014.10.005
25. Rivas-San Vicente M, Plasencia J. Salicylic acid beyond defence: Its role in plant growth and development. Journal of Experimental Botany. 2011; doi: 10.1093/jxb/err031
26. Morita-Yamamuro C, et al. The Arabidopsis gene CAD1 controls programmed cell death in the plant immune system and encodes a protein containing a MACPF domain. Plant Cell Physiol. 2005; doi: 10.1093/pcp/pci095
27. Han JY, In JG, Kwon YS, Choi YE. Regulation of ginsenoside and phytosterol biosynthesis by RNA interferences of squalene epoxidase gene in Panax ginseng. Phytochemistry. 2010; doi: 10.1016/j.phytochem.2009.09.031
28. Wang K, Senthil-Kumar M, Ryu C-M, Kang L, Mysore KS. Phytosterols Play a Key Role in Plant Innate Immunity against Bacterial Pathogens by Regulating Nutrient Efflux into the Apoplast. PLANT Physiol. 2012; doi: 10.1104/pp.111.189217
29. Gupta SK, Rai AK, Kanwar SS, Sharma TR. Comparative analysis of zinc finger proteins involved in plant disease resistance. PLoS One. 2012; doi: 10.1371/journal.pone.0042578
30. Soll J, Schleiff E. Protein import into chloroplasts. Nature Reviews Molecular Cell Biology. 2004; doi: 10.1038/nrm1333
31. Stief A, et al. Arabidopsis miR156 Regulates Tolerance to Recurring Environmental Stress through SPL Transcription Factors. Plant Cell. 2014; doi: 10.1105/tpc.114.123851
32. Michaels SD, Amasino RM. Memories of winter: vernalization and the competence to flower. Plant, Cell Environ. 2000; doi: 10.1046/j.1365-3040.2000.00643.x
33. Liu G, Holub EB, Alonso JM, Ecker JR, Fobert PR. An Arabidopsis NPR1-like gene, NPR4, is required for disease resistance. Plant J. 2005; doi: 10.1111/j.1365-313X.2004.02296.x
34. Gutterson N, Reuber TL. Regulation of disease resistance pathways by AP2/ERF transcription factors. Current Opinion in Plant Biology. 2004; doi: 10.1016/j.pbi.2004.04.007
35. Mitchell DA, Vasudevan A, Linder ME, Deschenes RJ. Protein palmitoylation by a family of DHHC protein S-acyltransferases. J Lipid Res. 2006; doi: 10.1194/jlr.R600007-JLR200
36. Li Y, Scott R, Doughty J, Grant M, Qi B. Protein S -Acyltransferase 14: A Specific Role for Palmitoylation in Leaf Senescence in *Arabidopsis*. Plant Physiol. 2016; doi: 10.1104/pp.15.00448
37. Sharmin S, et al. Xyloglucan endotransglycosylase/hydrolase genes from a susceptible and resistant jute species show opposite expression pattern following Macrophomina phaseolina infection. Commun Integr Biol. 2012; doi: 10.4161/cib.21422
38. Okazawa K, et al. Molecular cloning and cDNA sequencing of endoxyloglucan transferase, a novel class of glycosyltransferase that mediates molecular grafting between matrix polysaccharides in plant cell walls. J Biol Chem. 1993
39. Sakuma Y, et al. DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. Biochem Biophys Res Commun. 2002; doi: 10.1006/bbrc.2001.6299
40. Gkizi D, Santos-Rufo A, Rodríguez-Jurado D, Paplomatas EJ, Tjamos SE. The β-amylase genes: Negative regulators of disease resistance for Verticillium dahliae. Plant Pathol. 2015; doi: 10.1111/ppa.12360

Europe PMC Funders Author Manuscripts

41. Huibers RP, de Jong M, Dekter RW, Van den Ackerveken G. Disease-specific expression of host genes during downy mildew infection of *Arabidopsis*. Mol Plant Microbe Interact. 2009; doi: 10.1094/MPMI-22-9-1104
42. Carter C. The Vegetative Vacuole Proteome of *Arabidopsis thaliana* Reveals Predicted and Unexpected Proteins. PLANT CELL ONLINE. 2004; doi: 10.1105/tpc.104.027078
43. Castaño-Miquel L, et al. SUMOylation Inhibition Mediated by Disruption of SUMO E1-E2 Interactions Confers Plant Susceptibility to Necrotrophic Fungal Pathogens. Mol Plant. 2017; doi: 10.1016/j.molp.2017.01.007
44. Mur LAJ, Simpson C, Kumari A, Gupta AK, Gupta KJ. Moving nitrogen to the centre of plant defence against pathogens. Annals of Botany. 2017; doi: 10.1093/aob/mcw179
45. Gao Y, et al. Two Trichome Birefringence-Like Proteins Mediate Xylan Acetylation, Which Is Essential for Leaf Blight Resistance in Rice. Plant Physiol. 2017; doi: 10.1104/pp.16.01618
46. Slavov GT, et al. Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass Miscanthus sinensis. New Phytol. 2014; 201:1227–1239. [PubMed: 24308815]
47. Grinberg NF, et al. Implementation of Genomic Prediction in Lolium perenne (L.) Breeding Populations. Front Plant Sci. 2016; 7:1–10. [PubMed: 26858731]
48. Spindel J, et al. Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLoS Genet. 2015; doi: 10.1371/journal.pgen.1004982
49. Biazzi E, et al. Genome-wide association mapping and genomic selection for alfalfa (Medicago sativa) forage quality traits. PLoS One. 2017; 12:1–17.
50. Bian Y, Holland JB. Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity (Edinb). 2017; doi: 10.1038/hdy.2017.4
51. Resende RT, et al. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity (Edinb). 2017; doi: 10.1038/hdy.2017.37
52. Hayes BJ, Lewin HA, Goddard ME. The future of livestock breeding: Genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends in Genetics. 2013; doi: 10.1016/j.tig.2012.11.009
53. Pryce JE, Daetwyler HD. Designing dairy cattle breeding schemes under genomic selection: A review of international research. Animal Production Science. 2012; doi: 10.1071/AN11098
54. Wientjes YCJ, Veerkamp RF, Calus MPL. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics. 2013; doi: 10.1534/genetics.112.146290
55. Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012; doi: 10.1186/1297-9686-44-4
56. Pliura A, Vaidotas L, Vytautas S, Edmundas B. Performance of twenty four european Fraxinus excelsior populations in three lithuanian progeny trials with a special emphasis on resistance to Chalara fraxinea. Balt For. 2011
57. Gautier M, et al. Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. Mol Ecol. 2013; 22:3766–3779. [PubMed: 23730833]
58. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; doi: 10.1093/bioinformatics/btu170
59. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013
60. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; doi: 10.1093/bioinformatics/btp352
61. Kofler R, Pandey RV, Schlötterer C. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics. 2011; 27:3435–3436. [PubMed: 22025480]

Europe PMC Funders Author Manuscripts

62. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; doi: 10.1093/bioinformatics/btn129

63. Wei T, Simko V. Package 'corrplot: visualization of a correlation matrix' (v.0.84). 2017

64. Landis JR, Heyman ER, Koch GG. Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. Int Stat Rev / Rev Int Stat. 1978; doi: 10.2307/1402373

65. Storey JD, Bass AJ, Dabney A, Robinson D, Warnes G. qvalue: Q-value estimation for false discovery rate control. R. 2019

66. Laetsch DR, Blaxter ML, Leggett RM. BlobTools: Interrogation of genome assemblies [version 1; referees: 2 approved with reservations]. F1000Research 2017. 2017

67. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin). 2012; doi: 10.4161/fly.19695

68. Waterhouse A, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018; doi: 10.1093/nar/gky427

69. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015; doi: 10.1038/nprot.2015.053

70. Schrödinger L. The PyMOL molecular graphics system, version 1.8. 2015

71. Trott O, Olson AJ. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. J Comput Chem. 2010; doi: 10.1002/jcc

72. Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. Methods Mol Biol. 2015; doi: 10.1007/978-1-4939-22697_19

73. Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J. 2011; doi: 10.3835/plantgenome2011.08.0024

74. Endelman JB, Jannink J-L. Shrinkage Estimation of the Realized Relationship Matrix. Genes|Genomes|Genetics. 2012; doi: 10.1534/g3.112.004259

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

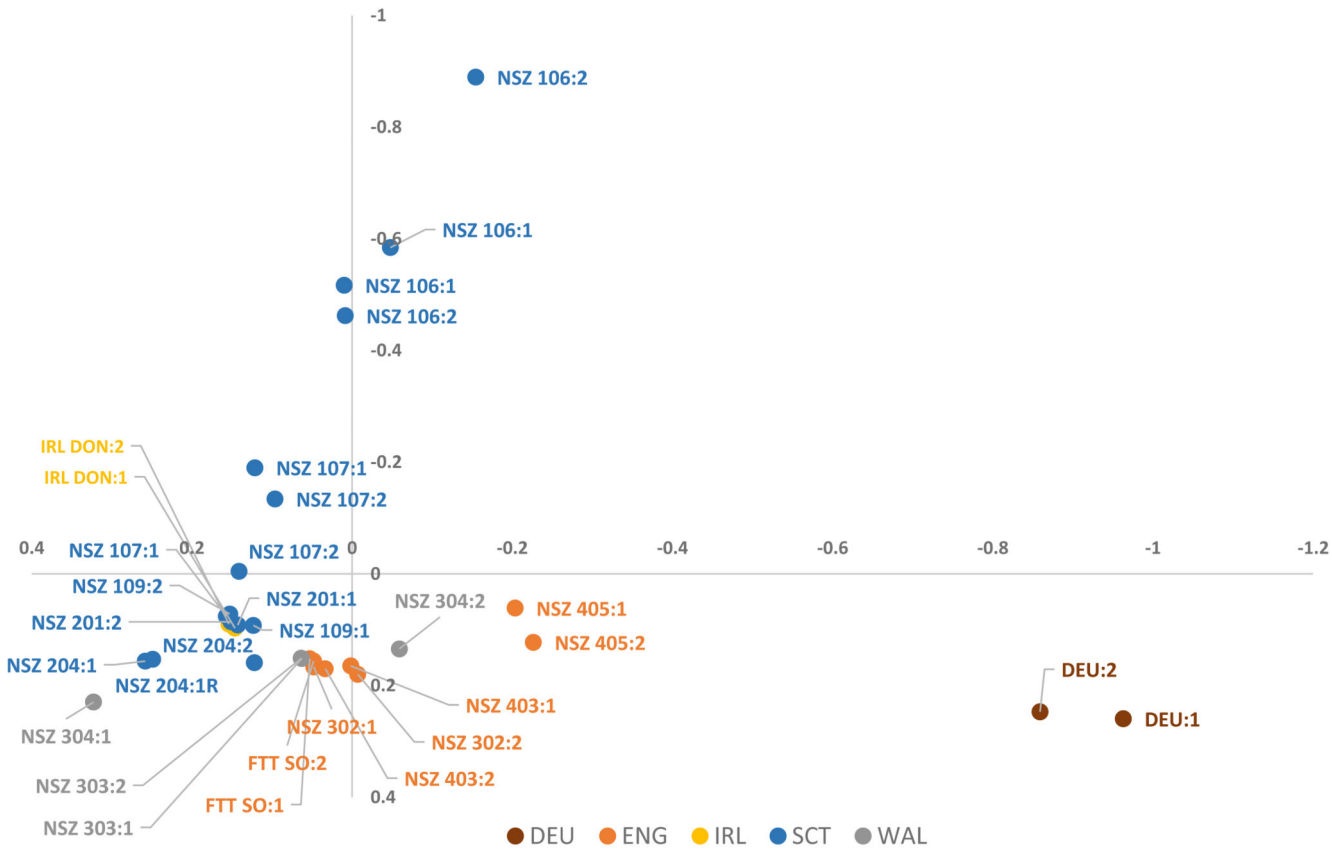


Figure 1. Summary of variation among the sequenced DNA pools using Correspondence Analysis (CA).

Major allele frequencies were used for all 31 seed source populations (including replicate). Numbers after seed source code correspond to health status (1 - healthy or 2 - infected by ADB). The vertical axis represents Principal Coordinate 1, which accounts for 10% of the variation and the horizontal axis represents Principal Coordinate 2, which accounts for 9% of the variation.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

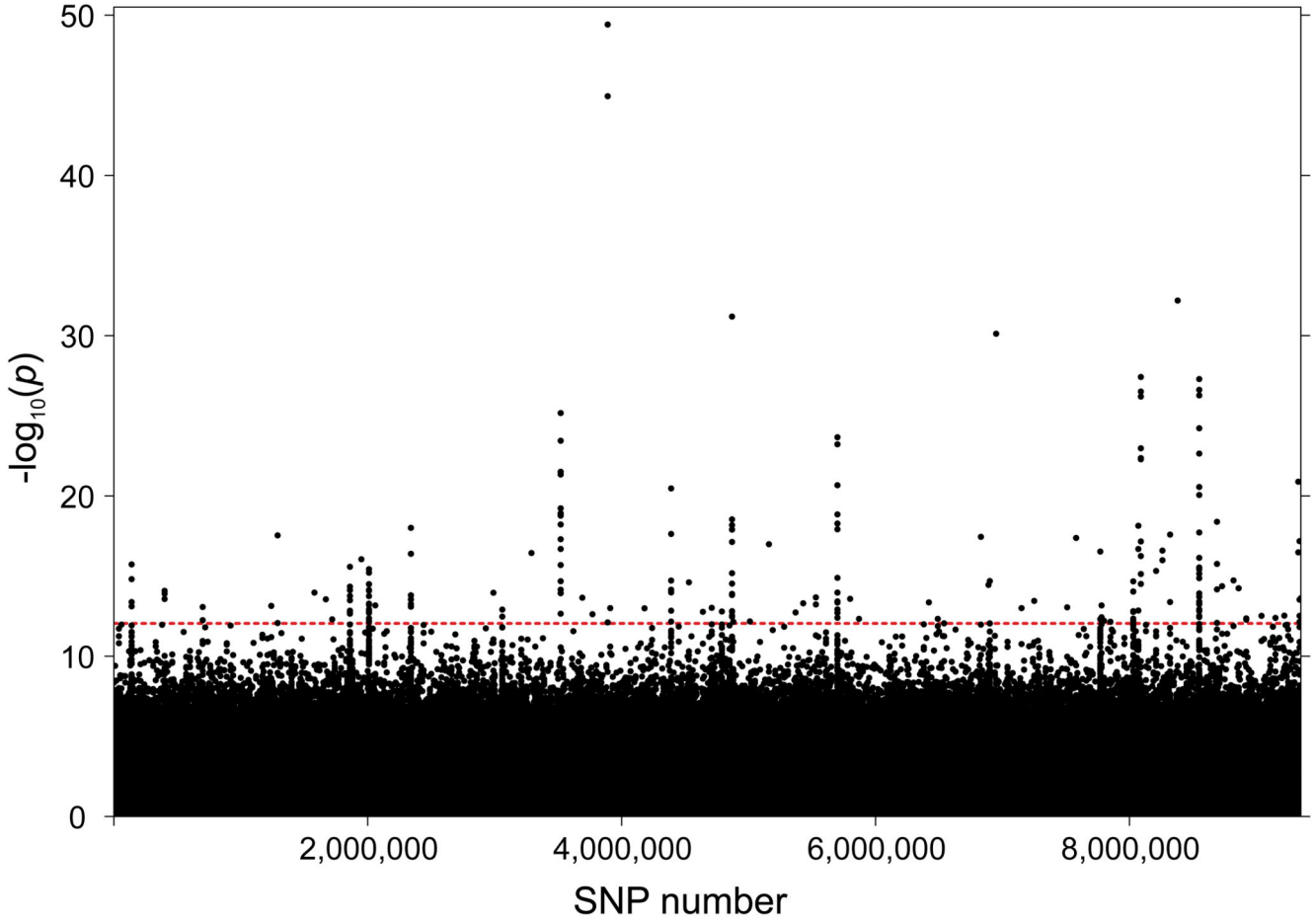


Figure 2. Manhattan plot for pool-seq genome-wide association study of tree health under natural ash dieback inoculation.

For each SNP a $-\log_{10}(p)$ value is shown. The green line represents the $p = 1 \times 10^{-13}$ threshold. Loci are ordered by position in the *F. excelsior* reference genome (BATG0.5).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

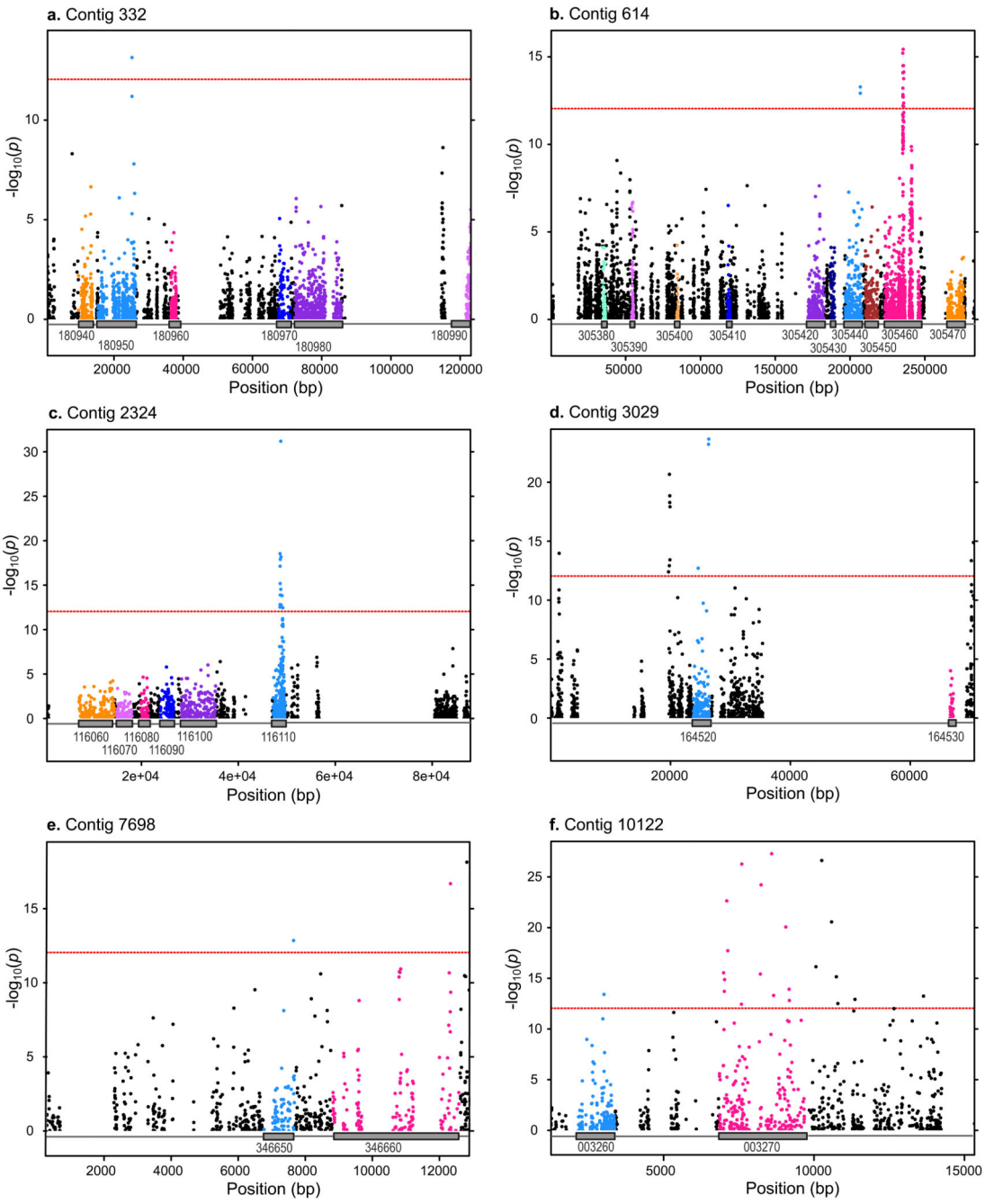


Figure 3. Manhattan plots for contigs containing genes with missense variants associated with tree health under natural ash dieback inoculations.

Points representing SNPs within genes are colored and those genes containing missense SNPs are named above the plot in the same colour as the points representing SNPs within them. The red line represents the $p = 1 \times 10^{-13}$ threshold.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

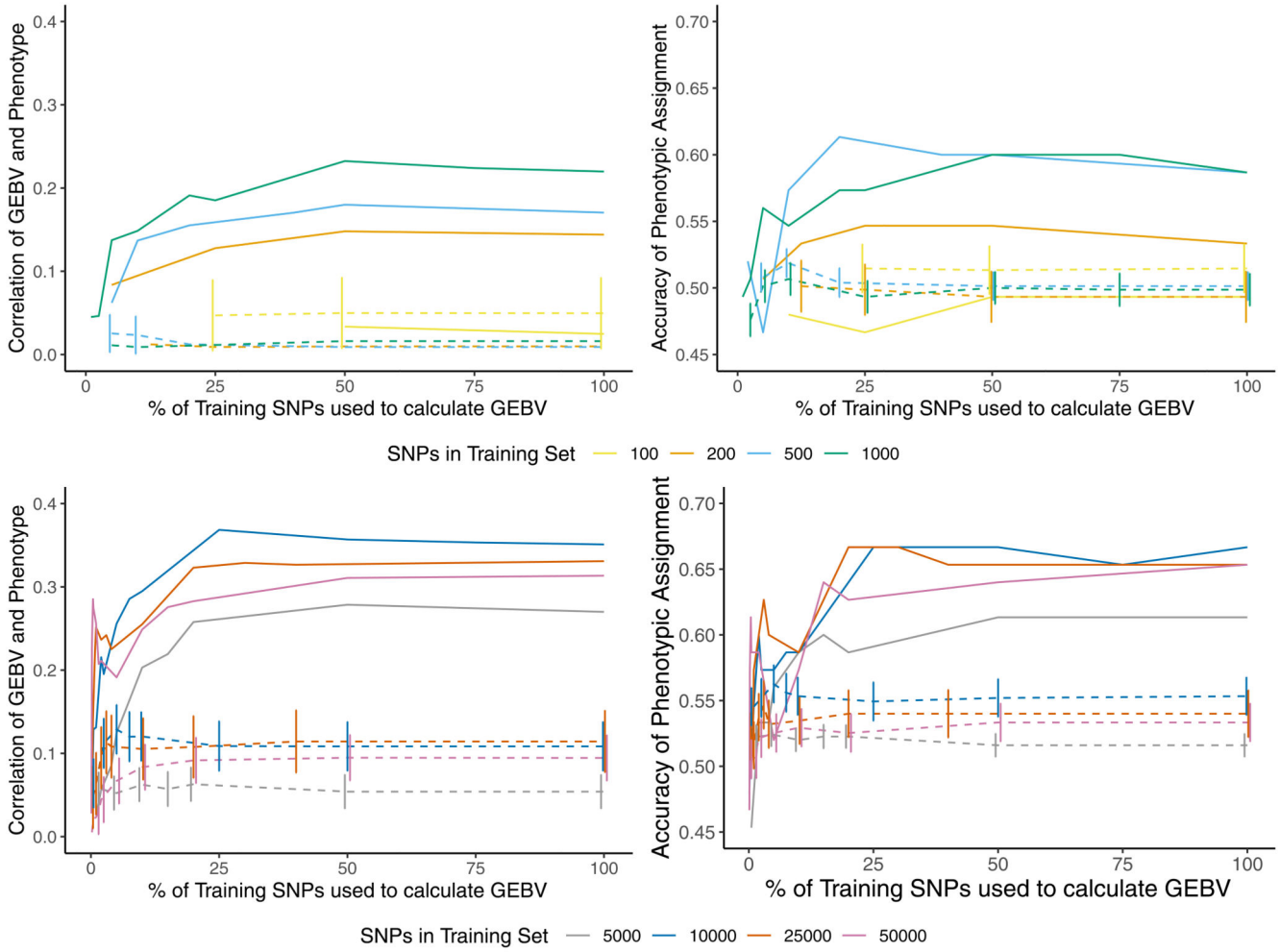


Figure 4. Performance of genomic prediction models for health under ash dieback pressure.

For 150 individual ash trees, with models trained on pooled sequencing of 1250 trees, using varying numbers of SNPs in training and test sets. Solid lines show results for SNPs selected using the pool-seq GWAS; dashed lines show mean results for repeated runs (n=10) of randomly selected SNPs, with bars indicating standard error. Left column: correlation of genomic estimated breeding value (GEBV) with observed health status. Right column: accuracy of health status assignment from GEBV.

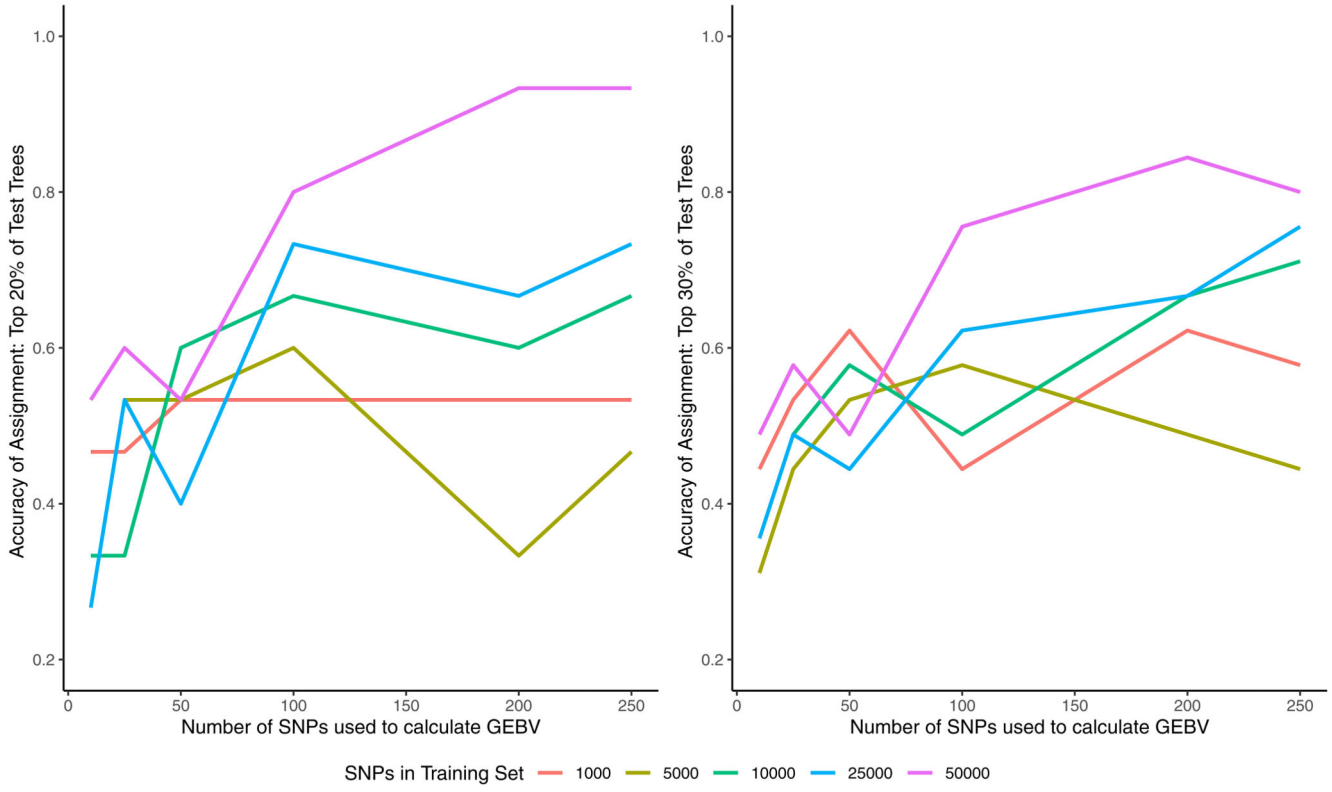


Figure 5. Performance of genomic prediction models for selection.

Genomic prediction accuracy of assignment of health status for the (left) top 20% and (right) top 30% of test population trees by GEBV, using 1000 to 50,000 SNPs identified by GWAS in the training set and use of ten to 250 SNPs in the testing set.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Nat Ecol Evol. Author manuscript; available in PMC 2020 May 18.

Table 1

Ash genes likely to be affected by top GWAS candidate SNPs.

Based on the top 192 hits by p-value (with $-\log_{10}(p) > 13$): (1) Genes that contain one or more significant SNP loci altering protein sequence; (2) Genes containing SNPs that are transcribed but not translated (synonymous changes, and changes in UTRs and introns); (3) Genes that are within 5Kb of significant SNP loci and the closest gene to those loci; Asterisks (*) mark genes that have evidence for involvement with disease resistance in other species. The "Gene" column gives the final six digits for the full gene names for the annotation of the ash genome[2], which are in the form FRAEX38873_v2_000######. Details of amino acid changes in missense variants can be found in Table S5.

Contig location	Gene	Predicted function	Variant functions (positions)
1) Genes containing SNPs that affect protein sequence			
Contig10122: 2113-3371	003260*	BED finger-NBS-LRR resistance protein (for model see Extended Data Fig. 5a)	1x missense variant (3018)
Contig10122: 6838-9688	003270*	Protein CPR-5-like (LOC111390874), transcript variant X1, mRNA	5x 3' UTR variant (6993, 7018, 7026, 7098, 7133) 2x 5' UTR premature start codon gain variant (9062, 9181) 2x 5' UTR variant (8656, 9172) 3x intron variant (8216, 8239, 8592) 7x upstream gene variant (10064, 10257, 10742, 10793, 11362, 12658, 13638) 1x missense variant (7599)
Contig2324: 47147-49656	116110	60S ribosomal protein L4-1 (LOC111391733), mRNA (for model see Extended Data Fig. 5d)	4x missense variant (48646, 48672, 48665, 48775) 9x synonymous variant (48620, 48623, 48626, 48629, 48764, 48809, 48827, 49028, 49180)
Contig3029: 23834-26488	164520	F-box/kelch-repeat protein SKIP6 (LOC111408673), mRNA (for model see Extended Data Fig. 5b)	1x 5' UTR variant (26379) 7x downstream gene variant (19676, 19824, 19878, 19882, 19907, 19808, 19921) 1x missense variant (26333)
Contig332: 15436-26198	180950	Protein DAMAGED DNA-BINDING (for model see Extended Data Fig. 5c)	1x missense variant (25205)
Contig614: 196876-208043	305440*	Uncharacterized LOC111377332 (LOC111377332), transcript variant X1, mRNA	1x missense variant (206888) 1x synonymous variant (206889)
Contig7698: 8815-12615	346660	Protein HEAT INTOLERANT 4-like (LOC111409690), mRNA(1)	1x missense variant (12331) 1x upstream gene variant (12819)
2) Genes containing SNPs that are transcribed but not translated			
Contig2329: 13133-19211	116430	Uncharacterized LOC111374226 (LOC111374226), transcript variant X2, mRNA	1x synonymous variant (13617)
Contig2747: 43908-51835	145630	VIN3-like protein 1 (LOC111390514), transcript variant X2, mRNA	1x synonymous variant (45617)
Contig4397: 39490-43181	234590*	WPP domain-interacting protein 1-like (LOC111407140), mRNA	1x synonymous variant (42379)

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Contig location	Gene	Predicted function	Variant functions (positions)
Contig1096: 98748-106855	013250*	MACPF domain-containing protein CAD1-like (LOC111379406), mRNA	1x 3' UTR variant (98777) 1x intron variant (103716)
Contig1454: 115669-119933	047060	Short-chain dehydrogenase TIC 32, chloroplastic-like (LOC111372928), transcript variant X2, mRNA	1x intron variant (118417)
Contig1506: 8961-15605	051390	Probable boron transporter	1x intron variant (13054)
Contig1589: 87297-124585	057960	Beta-taxilin (LOC111407559)	1x intron variant (123113)
Contig1795: 173977-176065	074310	Squamosa promoter-binding-like protein 8 (LOC111383449), mRNA	1x 3' UTR variant (174158)
Contig2034: 21612-30725	094440	Regulatory-associated protein of TOR 1 (LOC111407995), mRNA	1x 3' UTR variant (29838)
Contig2185: 59512-60735	105920	Uncharacterized LOC111409367 (LOC111409367), mRNA	1x 5' UTR variant (60622)
Contig23: 363875-372515	114040	ATP synthase subunit O, mitochondrial-like (LOC111411675), mRNA	1x intron variant (371764) 3x upstream gene variant (373188, 373372, 375563)
Contig2870: 78472-84675	154470	Uncharacterized LOC111404100	2x intron variant (83904, 83931)
Contig31173: 6556-7507	168770	Protein LATE FLOWERING-like (LOC111406993), mRNA	1x 5' UTR variant (6633)
Contig3809: 43625-54185	207550	receptor-like cytosolic	1x intron variant (50024)
Contig3889: 1-4475	211580*	Squalene monooxygenase-like (LOC111410179), mRNA	1x intron variant (957)
Contig4494: 40325-50726	238810	Uncharacterized LOC111381639	1x 3' UTR variant (40482)
Contig5196: 685-5813	266510	Zinc finger CCCH domain-containing protein 11-like (LOC111366362), transcript variant X3, mRNA	1x intron variant (3930)
Contig614: 223287-246926	305460*	Protein PHR1-LIKE 3-like (LOC111377335), mRNA	14x intron variant (235226, 235272, 235318, 235327, 235343, 235356, 235367, 235506, 235514, 235705, 235801, 235831, 235852, 235915)
Contig6272: 67196-77814	308800	Probable DNA helicase MCM8 (LOC111365493), transcript variant X2, mRNA	2x intron variant (70778, 71059)
Contig6641: 5399-9018	319390	Uncharacterized LOC111408674 (LOC111408674), mRNA	1x intron variant (7471)
Contig754: 35704-44204	342270	Protein LIKE COV 2-like (LOC111397136), mRNA	2x intron variant (38068, 42193)
Contig754: 75965-84594	342280	Uncharacterized LOC111408663 (LOC111408663), transcript variant X5, misc_RNA	1x 5' UTR variant (76154)

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Contig location	Gene	Predicted function	Variant functions (positions)
Contig7698: 6766-7686	346650	Pentatricopeptide repeat-containing protein At4g39620, chloroplastic-like (LOC111408678), transcript variant X2, mRNA	1x 3' UTR variant (7650)
Contig87: 379742-385962	372350	Uncharacterized LOC111393674 (LOC111393674), mRNA	3x intron variant (383677, 383731, 383732)
Contig8942: 7696-39701	378970	Uncharacterized LOC111377872 (LOC111377872), transcript variant X8, mRNA	1x intron variant (20201)
3) Genes within 5Kb upstream or downstream from candidate SNPs			
Contig1224: 126407-127675	025560*	Probable xyloglucan endotransglucosylase/hydrolase protein 28 (LOC111399252), mRNA[(3)]	1x upstream gene variant (130319)
Contig1607: 24206-51892	059350	Low affinity sulfate	1x upstream gene variant (22510)
Contig16137: 733-2920	059880	60S Ribosomal protein L30-like (LOC111409078), transcript variant X1, mRNA	1x upstream gene variant (3808)
Contig168: 17488-20603	065110	E3 ubiquitin-protein ligase RNF170-like (LOC111409836), transcript variant X3, mRNA	2x upstream gene variant (16009, 16035)
Contig1931: 125622-128117	086130	Oleoyl-acyl carrier protein thioesterase 1, chloroplastic-like (LOC111385815), mRNA[(1)]	2x downstream gene variant (130264, 130505)
Contig2441: 6066-10198	124500	Ent-kaurene oxidase, chloroplastic-like (LOC111394477), mRNA	1x upstream gene variant (5627)
Contig3029: 66605-67270	164530	Uncharacterized LOC111408676 (LOC111408676), transcript variant X3, mRNA	1x upstream gene variant (70203)
Contig349: 148156-150262	190500*	Ethylene-responsive transcription factor ERF098-like (LOC111379140), mRNA[(1)]	2x downstream gene variant (152443, 152551)
Contig3945: 23862-30945	214510	Basic Helix loop helix protein A (LOC111388546) mRNA	1x upstream gene variant (32110)
Contig4503: 35427-41965	239330	Vacuolar protein sorting-associated protein 20 homolog 2-like (LOC111393567), mRNA	1x upstream gene variant (44192) 2x intergenic region (48262, 48540)
Contig454: 79796-107635	241210	Kinesin-like protein KIN-7K, chloroplastic (LOC111375100), mRNA	1x upstream gene variant (108478)
Contig490: 155143-163521	255180	Casein kinase 1-like protein HD16 (LOC111366886), mRNA	1x upstream gene variant (152485)
Contig4981: 40436-41111	258470*	F-box/FBD/LRR-repeat protein At1g13570-like (LOC111367195), transcript variant X2, mRNA	1x upstream gene variant (39661)
Contig508: 82928-91585	262070	Putative zinc transporter At3g08650 (LOC111388858), mRNA	1x downstream gene variant (94609)
Contig558: 91578-97432	282910	Nitrate regulatory gene2 protein-like (LOC111409481), mRNA	1x upstream gene variant (101905)

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Contig location	Gene	Predicted function	Variant functions (positions)
Contig558: 116946-118620	282920	Uncharacterized LOC111409076 (LOC111409076), mRNA	2x downstream gene variant (119168, 119172) 1x upstream gene variant (114672)
Contig558: 139766-144371	282930	Uncharacterized LOC111409077 (LOC111409077), transcript variant X3, mRNA	1x upstream gene variant (138012)
Contig592: 225074-229835	296810*	Ankyrin repeat-containing protein NPR4-like (LOC111379708), mRNA	1x downstream gene variant (222554)
Contig6316: 100-2973	310310	Calmodulin-binding protein 60 A-like (LOC111368134), transcript variant X3, mRNA	2x upstream gene variant (4636, 4779)
Contig7472: 22326-25723	340820*	Dehydration-responsive element-binding protein 2C-like (LOC111397561), transcript variant X1, mRNA	5x upstream gene variant (17466, 17478, 17533, 18485, 18547)
Contig754: 16146-18115	342250	Ethylene-responsive transcription factor ERF113-like (LOC111408666), mRNA	1x upstream gene variant (15367)
Contig754: 19567-26425	342260*	Protein S-acyltransferase 8-like (LOC111408665), mRNA	2x upstream gene variant (28606, 28760)
Contig8383: 2536-8425	364260	Pentatricopeptide repeat-containing protein At4g39620, chloroplastic-like (LOC111408678), transcript variant X2, mRNA	1x upstream gene variant (9024)

[1]Blast performed using cDNA sequences