Genomic evidence of natural selection for ash dieback resistance

Carey L. Metheringham[1,2], William J. Plumb[1,2,3], Jonathan J. Stocks[1,2], Laura J. Kelly[1,2], Miguel Nemesio Gorriz[3], Justin Moat[2], Richard J. A. Buggs[1,2], Richard A. Nichols[1]

[1] School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK.

[2] Royal Botanic Gardens, Kew, Richmond upon Thames, UK.

[3] Forestry Development Department, Teagasc, Dublin, Republic of Ireland

Correspondence: r.a.nichols@qmul.ac.uk; r.buggs@kew.org

Large changes in quantitative traits through small shifts in allele frequencies at many genetic loci driven by natural selection lie at the heart of the neo-Darwinian synthesis[1,2] but are hard to fully demonstrate in the wild[3–6]. The spread of a natural pathogen into a new host population containing standing genetic variation for quantitative resistance provides an opportunity to demonstrate such shifts occurring in real time[7–10]. The invasive fungal pathogen (*Hymenoscyphus fraxineus*) causes ash dieback disease in European ash (*Fraxinus excelsior)* populations. The disease is frequently lethal[11,12], but larger trees persist longer[11,13], often with reduced reproductive output[13]. Here, using whole genome resequencing of an infected natural ash tree population in England, we measure allele frequencies at 7985 single nucleotide polymorphisms (SNPs) previously associated with resistance to the pathogen[14]. The average genetic merit for resistance of trees established after the start of the epidemic, calculated as Genomic Estimated Breeding Values (GEBV) from the 7985 loci, was higher than that of related adults from the pre-epidemic generation. We estimate that, to produce a GEBV shift of this magnitude, would require truncation selection eliminating at least 13% of the juvenile population. Even this rapid rate of change under natural selection may be insufficient for evolutionary rescue[15] of European ash populations, but change could be accelerated by a breeding programme informed by genomic selection. To this end, we show that for predicting the health of their offspring, the GEBVs of parent trees perform better than a phenotypic assessment of the ash dieback damage parent trees have suffered.

Our study site, Marden Park wood, is an ancient semi-natural woodland dominated by *F. excelsior* where the pathogen *H. fraxineus* is thought to have been present since 2012. Phenotypic assessments of this woodland in 2019, when we collected ash tissue samples for sequencing, found *H. fraxineus* symptoms on the majority of trees. Damage from the fungus had further increased by 2021 especially in juvenile trees (Figure 1). This fits with widespread documentation of the progress of the ash dieback epidemic throughout Europe[11,12].

We generated short read sequence data for 580 individuals (128 pre-epidemic adults and 452 post-epidemic juveniles) and 30 technical replicates at approximately 11X whole genome coverage and from these data estimated allele frequencies at SNP loci. In a previous study, based on 38,784 ash trees from diverse provenances growing in ash dieback field trials, we had sequenced the 623 least-damaged trees and 627 highly damaged trees and identified 10,000 SNP loci associated with these phenotypes[14]. In that study these SNPs were used to train genomic prediction models to estimate the Genomic Estimated Breeding Values (GEBVs) of individual trees. The 10,000 effect sizes used to calculate the GEBVs were validated by being used to successfully predict the resilience of trees in a test population[14]. Of these 10,000 SNPs, 7,985 were variable in the Marden Park dataset and passed allele frequency and quality thresholds. This smaller number of polymorphic SNPs reflects the lower genetic diversity present in Marden Park wood than in the planted trials of trees chosen from diverse accessions. Of the 2,015 SNPs that were not variable in the Marden Park population, 1,055 were fixed for the allele associated with low ash dieback damage in the planted trials and 960 for the allele associated with high damage.

We calculated GEBVs for the Marden Park trees from the 7,985 polymorphic SNPs using the parameters of the genomic prediction model trained on the field trials[14]. A visual assessment of ash dieback damage showed only a weak relationship with GEBV scores (Extended Data Figure 3 and 4), possibly because of the different phenotype used in adults (canopy cover) and the reduced control of age and environment for juveniles, compared to the field trials.

Rather than using correlations with phenotypes, which may be only indirectly related to survival, we sought more direct evidence for the action of natural selection by comparing the GEBV scores of adults and juveniles. There was a shift in mean GEBV in the predicted direction: the mean GEBV of juvenile trees ($\mu = 0.22$, $n = 452$) was higher than the mean GEBV of adult trees

(μ= 0.15, n = 128) (Figure 2A). The statistical significance of this increase in score cannot be evaluated by treating each juvenile tree as an independent observation, since some adults have left multiple offspring. Instead, we compared the GEBV score of each individual with that predicted from its ancestry. If the surviving juveniles have atypically high GEBV scores, then on average their scores should be higher than predicted from their ancestry. To capture this ancestry, we characterised relatedness between juveniles and the adults by selecting 5,793 loci that were not in close linkage with the 7,985 GEBV sites or with one another, had no missing data and had minor allele frequencies of over 0.3. We expected these loci to be near-neutral in their effect on ash dieback susceptibility but able to predict the GEBV in juveniles on the basis of relatedness alone. By way of illustration, genomic prediction using these 5,793 loci trained on the juvenile GEBV scores produced a perfect correlation with juvenile GEBVs (r=1, Extended Data Figure 5). To evaluate the change in GEBV across generations, the model was trained on the adult trees and used to predict the juvenile scores. These predictions were significantly lower (Pr(>F) = 0.0004) than the actual juvenile GEBVs (regression in Figure 2B). Thus, juvenile trees had higher GEBVs than predicted on their relationship to the adult trees, in the absence of selection. The observed shift is 0.058, slightly lower than the raw estimate of 0.07 from the difference of adult and parent means; it is 15% of the standard deviation of GEBV distribution in the juvenile trees. To characterise the intensity of selection required to produce this shift, we can calculate the level of truncation selection that would produce such a shift in GEBV. If we assume that the narrow sense heritability of ADB susceptibility is 0.4, as previously estimated by several field-trial studies[16], it would require selection eliminating the most susceptible 13% of the juvenile population. This level of selection appears plausible, given the 85% mortality seen in some mature ash plantations due to ash dieback over a 20 year period[11,12], of which only a small portion would need to be conditional on the GEBV score to have produced the response that we have detected. If heritability in the woodland population is lower than the field-trial estimates, due to greater environmental variability and lower genetic diversity, the estimate of selection would be correspondingly larger. For heritabilities of 0.2 or 0.1, the estimates would be 29% or 60% elimination respectively.

If, as we propose, this shift in GEBV is due to selection, then we would predict a pattern in which the change in allele frequency depends on the effect size and the frequency of the beneficial allele. We detected this pattern (Figure 3) in a regression of allele frequency change on effect-size x p(1-p), where p is the minor-allele frequency. The relationship was highly significant (r = 0.0015, p = 0.00038, F = 12.64 on 1 and 7983 DF). The loci previously

estimated to be of large effect (the largest 25%) also showed a stronger association with our visual assessments of dieback phenotype than randomly selected loci in both adults and juveniles ($P < 10^{-14}$, combined over 100 random selections using Fisher's method).

A second test of our interpretation of the shift in GEBV can be conducted in the case of the 121 juveniles for which we were able to assign at least one parent with confidence (using the sequoia R package[17] on 1000 SNPs). In these cases, we could predict the allele frequencies expected by random transmission. The observed frequencies were significantly different from these random expectations ($q < 0.05$, Wilcoxon rank sum test with Bonferroni correction). The advantageous alleles, associated with low susceptibility to ash dieback, were over-represented, the excess frequency being correlated with the effect size ($r = 0.0595$, $p < 10^{-6}$). This pattern again suggests that the offspring carrying these alleles had a selective advantage, making them more likely to survive to become part of our sample (compared to siblings that did not).

Thus, we have documented a heritable micro-evolutionary adaptive change occurring over a single generation. This selection may have occurred partly by reduced seed or pollen production from adult trees damaged by ash dieback[13,18] and partly through rapid death of young trees infected by ash dieback, constituting a missing fraction that did not survive. To have demonstrated such a shift in heritable variation without genome-sequencing would have involved producing large numbers of clones or offspring from many adult and juvenile trees in the woodland and growing these in uniform conditions for several years under ash dieback inoculum pressure. Our approach was possible because we could build on the estimates of the effect size at 7,985 loci obtained from large field trials involving 38,784 individuals raised in standardised conditions. The direction and magnitude of the observed changes in allele frequency correspond to these effect size estimates, showing that they are broadly applicable in the semi-natural conditions of Marden Wood.

Few other studies have thus far sought evidence for micro-evolutionary change via shifts in GEBVs, though several studies in animals have examined shifts in predicted breeding values based on pedigrees[19]. A recent study showed a shift in GEBVs for adult weight in a 35 year data set for Soay sheep, though phenotypic body weight decreased over this time span[20]. Genomic prediction approaches have a wide future applicability in evolutionary studies[21].

The action of natural selection in the wild for increased health of European ash under ash dieback is good news, but whether the rate of change will be sufficient for evolutionary rescue to occur[15], and whether its end point will be a fully resistant tree, is as yet unknown. The population we studied is fixed for alleles associated with high susceptibility at more than 900 loci; at these loci gene flow from other populations will be essential to introduce beneficial alleles. Our parentage analyses suggest ongoing migration of pollen and/or seed into the population, but this may decrease as ash becomes more scarce in the landscape. As low susceptibility is highly polygenic, the speed of evolution may be inhibited by linkage among loci[22,23]. As the epidemic progresses, selective pressure may decrease as the pathogen follows the host demographics to become more scarce in the landscape. This will diminish the speed of evolution, especially given that heritability may be low in highly variable woodland environments. In many British ash woodlands, natural regeneration of ash is scarce due to deer browsing, so deer management will be essential for natural selection to be effective. Further interventions could enhance the rate of evolutionary change, for example, by introducing genetic diversity from other seed sources or from trees selected, or bred, for high GEBV.

Several breeding programmes for ash dieback resistance are currently underway in Europe, with parental trees selected based on their ash dieback damage phenotypes. In examining the efficacy of this approach, for the 121 juveniles for which we could identify parents, we found no significant correlation between our health phenotypes of the parent trees and the mean phenotypic health score of their likely offspring ($r = 0.112$, $p = 0.54$, $n = 37$, Figure 4). In contrast, we found significant correlation between the GEBV of the parent trees and the health score of their offspring ($r = 0.368$, $p = 0.029$, $n = 37$, Figure 4). This provides further evidence that our GEBVs of adult trees are better assessments of their genetic value for resistance to ash dieback than our leaf canopy phenotypes. A breeding programme for increased disease resistance based on GEBVs would allow more accurate selection of parents with high breeding values and would also allow parental pairs with complementary sets of alleles associated with resistance to be selected for crossing.

Methods

Site Selection

For this study we required an accessible ash-dominated woodland site in the south east of England, where *Hymenoscyphus fraxineus* has been present since 2012. We sought a site with the following characteristics: containing ash trees of a wide range of ages and sizes, including the presence of young naturally regenerated saplings; the older ash trees should be (as far as could be ascertained) naturally seeded rather than planted; the owners should have conducted few management interventions since 2012, and herbivore populations should be low enough to allow ongoing natural regeneration. We examined potential sites that fit these criteria, with Woodland Trust and Kent County Council. In August 2019 we selected a site within Stubbs Copse in Marden Park Wood in Surrey (51.2707, -0.0400) that has belonged to the Woodland Trust since 1994. This is on a chalk plateau of the North Downs at approximately 244m above sea level. It is in the Surrey Hills Area of Outstanding Natural Beauty and is a Site of Special Scientific Interest. It is an ancient semi-natural woodland site where mature and pole stage ash is dominant; with beech, whitebeam and oak as rarer components. It approximates to National Vegetation Classification (NVC) W12a - beech/oak/ash with a dog's mercury sub-community. When the Woodland Trust surveyed the site in 2019 they found that ash dieback was prevalent and only a "handful" of trees showed no signs of the disease with others showing "advanced stages of decline, especially in the younger, pole-staged trees". The 50 years plus management objectives for the wood are to increase its resilience to pests and diseases and maximise its biodiversity. The only management interventions planned from 2019-2024 are the removal of non-native species and the coppicing of some hazel in the understorey. Ash will only be felled where they pose a risk to site visitors and footpaths.

Sample and Phenotype Collection

From 16-20 September 2019 we surveyed and collected tissue samples from 784 ash trees in Stubbs Copse. We sought to exhaustively sample all ash trees, moving outwards from a central point. For each ash tree we recorded its size, level of ash dieback damage, and GPS coordinates. We labelled each tree with a plastic tag. We took a leaf sample from each tree, and when leaves were unavailable or inaccessible we took buds or twigs. Failing the availability or accessibility of any of these tissues, we took a cambial sample from the trunk using a hammer

203 and chisel, cutting out an area of bark approximately 2cm x 8cm. Sampled tissues were
204 immediately placed into plastic bags containing silica gel, to dry. For trees less than
205 approximately breast height, we measured tree size as height from ground level, and we scored
206 ash dieback damage using a scale similar to that of Pliura[24] with the difference that we had
207 fewer categories of dead trees. Trees we scored as 1 were completely dead, 2 were mostly
208 dead but with small sections of the main stem still living, 3 had infection which had spread into
209 the main stem, 4 had minor infections and 5 were healthy trees with no signs of infection in the
210 stem. For larger trees we measured size as diameter at breast height (DBH) and scored ash
211 dieback damage as an estimate of percentage leaf canopy loss from visual expectation from the
212 ground. On October 16 2019 and 21 and 22 January 2020 we re-collected tissue samples from
213 trees where the first DNA extraction attempt had failed, and as far as possible added aluminium
214 tags to the trees included in the study. On 4 and 12 March 2021 positions of trees were
215 collected using an RTK GPS (Bad Elf Pro), using NTRIP (Networked Transport of RTCM via
216 Internet Protocol) connection to the Imperial University unit (LICCOOGBRO), approximately 27
217 km from the site. On 18 and 18 August 2021, the trees were again re-scored for ash dieback
218 damage.
219
220 **Analysis of phenotypic and spatial data**
221
222 In 2021, 126 of the adult trees and 355 of the juvenile trees were re-phenotyped, and mapped in
223 detail (Extended Data Fig. 1). The phenotypes showed damage due to *H. fraxineus* in juvenile
224 trees increased between 2019 and 2021, with the mean health score decreasing from 3.6 (SE =
225 0.1, n = 642) in 2019 to 2.5 (SE = 0.1, n = 418) in 2021 (Figure 1); see methods section for
226 scoring system. The percentage of juvenile trees observed to be dead increased from 5.5% in
227 2019 to 40.9% in 2021. For adult trees, mean percentage canopy coverage was 22% (SE = 2.9,
228 n = 160) in 2019 and 20% (SE = 2.3, n = 126) in 2021 (Figure 1). The lower number of trees re-
229 phenotyped in 2021 was most likely due to loss of labels and accidental loss of juveniles due to
230 trampling or browsing. The 224 juvenile trees that were missing in 2021 had the same average
231 health score (3.6, SE = 0.1) in 2019 as all juvenile trees scored in 2019. Thus it seems unlikely
232 that their disappearance was mainly due to mortality caused by ash dieback. Moran's I (Moran,
233 1950) was calculated in order to estimate the degree of spatial auto-correlation within the
234 phenotypic data of trees which had their positions collected in 2021. The 2019 health scores of
235 the juvenile trees which were mapped in 2021 showed significant spatial auto-correlation
236 (Moran's I = 0.127, p = 0) but this was not the case in 2021 (Moran's I = 0.0135, p = 0.327). The

237 health of the adult trees did not show significant spatial auto-correlation in either year (2019:
238 Moran's I = 0.0546, p = 0.0538; 2021: Moran's I = 0.0452, p = 0.137). Spatial autocorrelation
239 among juvenile trees may be a product of local biotic and abiotic factors, or of relatedness
240 between juveniles in close proximity.
241
242 **DNA extraction and sequencing**
243
244 DNA was extracted from dried plant material using Qiagen DNeasy Plant Pro Kits or an adapted
245 CTAB extraction procedure[25]. DNA for 755 trees, including 52 replicates was sent to Novogene
246 for sequencing. Library preparation with NEBNext DNA Library Prep Kit and paired end Illumina
247 sequencing on NovaSeq 6000 was carried out by Novogene (Germany) with a total amount of
248 1.0μg DNA per sample. Genomic DNA was randomly fragmented to a size of 350bp by
249 shearing, then DNA fragments were end polished, A-tailed, and ligated with the NEBNext
250 adapter for Illumina sequencing, and further PCR enriched by P5 and indexed P7 oligos. PCR
251 products were purified using the AMPure XP system and resulting libraries were analysed for
252 size distribution by Agilent 2100 Bioanalyzer, quantified using real-time PCR, then sequenced.
253
254 **Sequence analysis**
255
256 We generated 0.71 to 19.92 Gb of read data per tree (μ = 9.93 Gb), giving 0.81 - 22.64 X
257 coverage (μ = 11.28) of the whole genome (assuming a genome size of 877Mbp[26]) for 649
258 individuals and 30 technical replicates. Reads containing adapters, more than 10%
259 undetermined bases or with more than 50% low quality (Qscore<= 5) bases were filtered out
260 with SAMtools (V1.9)[27] and remaining reads were aligned against the BATG0.5 *F. excelsior*
261 genome using the bwa mem function in the Burrows Wheeler Aligner (V0.7.17)[28] with minimum
262 seed length of 19, matching score of 1, mismatch penalty of 4, gap open penalty of 6 and gap
263 extension penalty of 1. For each sample, 3.0% to 98.9% of reads mapped to the *F. excelsior*
264 reference genome (μ = 86.3%). Samples with less than 50% of reads aligning to the reference
265 genome were excluded from further analysis.
266
267 SNP calling was performed using HaplotypeCaller, GenomicsDBImport and GenotypeGVCFs in
268 the Genome Analysis Toolkit (GATK V4.1.4.0)[29]. Sites with greater than 25% missing data were
269 removed and hard filtering was performed on remaining SNP variants using VariantFiltration in
270 GATK (V4.2.1.0) with the following parameters; read depth ≥ 10 and <200, variant confidence ≥

20, QD (variant confidence normalised by read depth) ≥ 4, mapping quality ≥ 40, FS ≤ 60 and minor allele frequency ≥ 0.01. After excluding samples with low alignment and filtering our read alignments for quality, coverage and indels, we called SNPs at 28,499,977 variable sites in 600 samples, including 25 replicates. Linkage disequilibrium pruning was performed using PLINK 1.9[30] with an independent pairwise r^2 threshold of 0.5, window size of 50 and step size of 5 using the option --indep-pairwise, resulting in a reduced set of 4,287,210 SNPs.

In a PCA of the SNP data performed in PLINK, the first and second principal components explained 12.1% and 9.2% of the genetic variance respectively (Extended Data Fig. 2). The identity of replicate samples was confirmed by calculating the genomic distance between samples using the –distance function in PLINK, with genomic distance recorded as 1 minus identity-by-state. The 25 technical replicate pairs had genomic distances of between 0.051 and 0.094 (μ = 0.063) (Supplementary Table 1) whereas all other samples (which likely include some trees with full sibling and parent-offspring relationships) had genomic distances of between 0.051 and 0.235 (μ = 0.214). The latter included 13 pairs of trees with genomic distances less than 0.1, which were likely to be close relatives (Supplementary Table 2). The genomic distance between our technical replicates suggests that we can expect 5-9% of genotypes to be miscalled, likely due to low sequence coverage. The sequence set with the highest coverage per replicate pair was used in all subsequent analyses.

Genomic Prediction of Susceptibility

A genetic estimated breeding value (GEBV) for each tree was calculated using effect sizes obtained from field trial data as described in Stocks (2019)[14]. Briefly, the trials comprised ash trees from 10 UK native seed zone (NSZ) provenances - including the zone in which Marden Park is found - together with one from Ireland and one from Germany. The trial trees were of similar age (~7 years) planted in randomised plots, with even spacing, on former agricultural land. For the genome-wide association study (GWAS) 38,784 trees were scored in the trials for ash dieback damage using a health assessment method very similar to the one we used on the juvenile trees in Marden Park wood. Only 792 fully healthy trees were found in the trials (1.96% of all trees), and of these 623 were used for GWAS and to train the GP model, together with 627 highly damaged trees, giving a total of 1250 trees selected from the extremes of ash dieback damage phenotypes[14]. An rrBLUP model was trained using pool-seq data from these 1250 trees, based on 10,000 loci that had the most significant association with disease susceptibility

in a pool-seq GWAS[14]. These 10,000 were able to predict the disease classification of individually sequenced test trees with 68% accuracy[14].

7,985 of these 10,000 "predictive" SNPs were variable within the Marden Park population. Effect sizes from this model, characterised as the additive effect of each copy of the alternate allele , were used to calculate GEBV of the Marden Park trees as $GEBV_j = \Sigma(g_j f_{ij}) + 1$, where $GEBV_j$ is the genomic estimated breeding value of individual j, $g_j$is the effect size at site i (Supp. Data 2 - from Stocks et al (2019)), and $f_{ij} \epsilon \{0,1,2\}$ is the frequency of the alternate alleles at site i in individual j (Supp. Data 3).

We assessed the relationship between GEBVs and tree health phenotypes by correlation of trees' GEBVs with the health scores of juvenile trees and the percentage canopy cover of the adult trees (Extended Data Fig. 3). The mean difference in health phenotypes of the trees in equal numbers of trees with the highest and lowest GEBVs were compared in increments of one tree and plotted in terms of percentage of trees included (Extended Data Fig. 4), finding a weak trend in the expected direction, which was not formally significant.

Detecting shifts in breeding value between adult and juvenile trees

We wished to investigate whether juvenile trees had higher GEBVs than we would have expected, based on the GEBVs of their adult relatives. All else being equal we would expect closely related individuals to have similar GEBV values. We should therefore be able to predict the GEBV of a tree from a sufficient number of SNPs that are unlinked to the GEBV sites and to each other, but which capture the pattern of relatedness. We therefore selected 5,793 such loci with a minor allele frequency of at least 0.3 (to ensure they were informative). To test whether that selection was sufficient to estimate GEBV on the basis of relatedness alone we fitted an rrBLUP(V4.6.1)[31] model on the juvenile trees as: $GEBV_J = \mu + X_J \beta_J + \varepsilon$, where $GEBV_j$ is the vector of breeding values of juvenile trees, μ is the mean, X_J is a matrix of juvenile genotypes ($\in$ {0,1,2}, giving the number of copies of the alternate allele) for the 5,793 SNPs, β_J is a vector of allelic effects (treated as normally distributed random effects) and the residual variance is $Var[\varepsilon]$. Effect sizes from this model were then used to calculate rGEBV, the prediction based on relatedness as: $rGEBV_J = \mu + \Sigma(g_j f_{ij})$, where g_j the effect size at site i, and f_{ij} is the genotype at site i in individual j. This overfitted model predicted GEBV perfectly (Extended Data Figure 5),

338 showing that this selection of SNPs are sufficient to predict GEBV, and could be used to make
339 predictions across generations (where over-fitting will not occur because the juveniles are not
340 included in the training dataset). To predict GEBV across generations we then trained the same
341 model on the adults, and then used the X_J of the juveniles to predict their GEBV values.
342 The level of truncation selection required to produce the observed shift in GEBV was found by
343 solving the equation $\Delta u = m(s, \sigma^2)$ for s, where m is the function giving the mean of the normal
344 distribution with variance σ^2 with the lowest area, s, removed and Δu is the required shift in
345 population mean given by the breeders equation: $0.15/h^2$. The values of σ^2 was set to $1/h^2$ to
346 correspond to the phenotypic variance unit additive genetic variance and $h^2=0.4$.
347
348 **Allelic shifts at causative loci**
349
350 If allele shifts are due to natural selection we would expect the magnitude of shifts to be related
351 to the effect size of the allele and the frequency of the allele in the population. The expected
352 change in allele frequency under selection, from f to f', is given by the standard equation $f' = (f^2$
353 $+ f(1-f)(1-s))/w$, where s is the selective effect and w the mean fitness.
354
355 The value, $\Delta f = f'-f$, is therefore proportional to $sf(1-f)$. Under the assumption that the selective
356 effect was proportional to the estimated effect sizes g_A, we plotted the observed change in
357 minor allele frequency, Δf, against $g_A f(1-f)$ in Figure 3. To visualise the trend more clearly we
358 grouped the loci into 200 bins (quantiles of $g_A f(1-f)$). The means of each bin are plotted in with
359 an area of each point being inversely proportional to the variance in Δf, to convey the relative
360 precision of the mean Δf estimate. The line is the fitted linear regression (carried out on the
361 individual Δf values).
362
363 To test if the sites identified as having large effects in the field trial[14] contributed to the visually
364 assessed dieback damage at Marden Park, a new genomic prediction was carried out using the
365 BGLR (V1.1)[32] package's implementation of the BayesB algorithm[33] (50,000 iterations with a
366 burn in of 2000). The genotypes comprised the sites in the top 25% (n=1996) of absolute effect
367 size, estimated from the field trial. These were combined with an equal number of control
368 genotypes randomly selected from across the genome (that were not part of the GEBV set).
369 This design was to discern any biological signal from the effect sizes that would be estimated
370 when fitting a model with a large number of loci to a relatively small number of observations.
371 The large effect loci had significantly larger estimated effect sizes in this new genomic prediction

372 model (characterised by their variance) across multiple selections of control loci. (Adults: Pr(>F)
373 < 1.1e-16, Juveniles: (Pr(>F) < 1.1e-16, combined over 100 random selections using Fisher's
374 method).
375
376 **Assignment of parentage**
377
378 Parentage assignment was performed with the sequoia R package (V2.3.1) [17], using a randomly
379 selected set of 1,000 SNPs having read depth > 20, minor allele frequency > 0.4, and an
380 estimated error rate < 0.01. Parentage assignment was run using the hermaphrodite "B" mode.
381 DBH was used to create a proxy for the birth year of adult trees (birth year = 100 - DBH) and
382 juvenile trees were assigned a birth year of 100. This allowed for larger, and therefore
383 presumably older, adult trees to be assigned as parents of younger adult trees, while
384 disallowing juveniles from being considered as parent trees. The maximum age of parents was
385 set as 99, allowing all adult trees to be considered as parents of the juveniles. The proxy years
386 were not intended to be an accurate estimate of tree age. Confidence of parentage assignment
387 was estimated within sequoia by simulating genotype data based on the estimated pedigree,
388 recalculating parentage assignment based upon simulated data and comparing the recalculated
389 pedigree to the original pedigree. The proportion of correct assignments over multiple runs of
390 the process gives an estimate of confidence in our parentage assignments.
391
392 Parentage was assigned to 153 trees, with a confidence value of 0.72. For 125 trees we could
393 assign a single parent and for 28 trees we could assign two parents. Of trees with assigned
394 parentage, 121 were classified as juveniles and 16 as adults. We identified 36 adult trees as
395 likely parents, with a mean of 4.9 offspring per parent tree. Twelve of the trees were only
396 assigned one likely offspring and 18 trees were assigned at least three offspring
397 (Supplementary Table 3). The largest set of offspring (n = 29) came from tree 51, which was
398 both large and centrally located (Extended Data Fig. 1), accounting for 21% of all trees with
399 assigned parentage. Of these 29 offspring, seven had tree 10 (located in close proximity to tree
400 51) as the other parent, making them full siblings.
401
402 For trees with assigned parents the expected frequency of the minor allele was calculated as;
403 $\frac{P1 + P2}{2}$, where $P1$ was the frequency of the allele in parent one and $P2$ was the frequency in
404 parent two. Where only one parent was identified the mean frequency of the minor allele across
405 the whole of the population was used as $P2$. The 7,985 sites of interest were tested for

significant deviation between expected and observed allele frequencies on a per site basis using a two sided Wilcoxon rank sum test with continuity correction, with p-values corrected for multiple testing using the Bonferroni correction. Correlation between effect size and the mean difference between observed and expected allele frequency in offspring at each site was tested using Pearson's product-moment correlation.

To test if parental GEBV was a better predictor of offspring health than parental phenotype, GEBV and percentage canopy cover of parent trees assigned by sequoia were each fitted against the mean health score of their likely offspring in linear models, weighted according to the number of likely offspring assigned to each parent tree (Figure 4). The fit of each model was summarised as the Pearson's correlation coefficient.

Data Availability

Raw reads for all sequenced samples are available on the ENA short read archive under accession PRJEB44697 (ERP128769). R scripts are available at https://github.com/CareyMetheringham/MardenPark

References

1. Fisher, R. A. *The genetical theory of natural selection*. (The Clarendon Press, 1930).
2. Falconer, D. S. *Introduction to Quantitative Genetics*. (Oliver and Boyd, 1962).
3. Pujol, B. *et al.* The Missing Response to Selection in the Wild. *Trends Ecol. Evol.* **33**, 337–346 (2018).
4. Endler, J. A. *Natural Selection in the Wild*. (Princeton University Press, 1986).
5. Merilä, J., Sheldon, B. C. & Kruuk, L. E. Explaining stasis: microevolutionary studies in natural populations. *Genetica* **112-113**, 199–222 (2001).
6. Charmantier, A., Garant, D. & Kruuk, L. E. B. *Quantitative Genetics in the Wild*. (OUP Oxford, 2014).
7. Hermisson, J. & Pennings, P. S. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005).
8. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).

439 9. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nat. Rev.*
440 *Genet.* **3**, 11–21 (2002).
441 10. Höllinger, I., Pennings, P. S. & Hermisson, J. Polygenic adaptation: From sweeps to subtle
442 frequency shifts. *PLoS Genet.* **15**, e1008035 (2019).
443 11. Coker, T. L. R. *et al.* Estimating mortality rates of European ash (*Fraxinus excelsior*) under
444 the ash dieback (*Hymenoscyphus fraxineus*) epidemic. *Plants, People, Planet* **1**, 48–58
445 (2019).
446 12. George, J.-P., Sanders, T. G. M., Timmermann, V., Potočić, N. & Lang, M. European-wide
447 forest monitoring substantiate the neccessity for a joint conservation strategy to rescue
448 European ash species (*Fraxinus* spp.). *Sci. Rep.* **12**, 4764 (2022).
449 13. Semizer-Cuming, D., Finkeldey, R., Nielsen, L. R. & Kjær, E. D. Negative correlation
450 between ash dieback susceptibility and reproductive success: good news for European ash
451 forests. *Ann. For. Sci.* **76**, 1–9 (2019).
452 14. Stocks, J. J., Metheringham, C. L., Plumb, W. & Lee, S. J. Genomic basis of European ash
453 tree resistance to ash dieback fungus. *Nature Ecology & Evolution* **3**, 1686–1696 (2019).
454 15. Carlson, S. M., Cunningham, C. J. & Westley, P. A. H. Evolutionary rescue in a changing
455 world. *Trends Ecol. Evol.* **29**, 521–530 (2014).
456 16. Plumb, W. J. *et al.* The viability of a breeding programme for ash in the British Isles in the
457 face of ash dieback. *Plants People Planet* **58**, 227 (2019).
458 17. Huisman, J. Pedigree reconstruction from SNP data: parentage assignment, sibship
459 clustering and beyond. *Mol. Ecol. Resour.* **17**, 1009–1024 (2017).
460 18. Semizer-Cuming, D., Chybicki, I. J., Finkeldey, R. & Kjær, E. D. Gene flow and reproductive
461 success in ash (Fraxinus excelsior L.) in the face of ash dieback: restoration and
462 conservation. *Ann. For. Sci.* **78**, 1–15 (2021).
463 19. Postma, E. Implications of the difference between true and predicted breeding values for
464 the study of natural selection and micro-evolution. *J. Evol. Biol.* **19**, 309–320 (2006).
465 20. Hunter, D. C. *et al.* Using genomic prediction to detect microevolutionary change of a
466 quantitative trait. *Proc. Biol. Sci.* **289**, 20220330 (2022).
467 21. McGaugh, S. E., Lorenz, A. J. & Flagel, L. E. The utility of genomic prediction models in
468 evolutionary genetics. *Proc. Biol. Sci.* **288**, 20210693 (2021).
469 22. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.*
470 **8**, 269–294 (1966).
471 23. Comeron, J. M., Williford, A. & Kliman, R. M. The Hill-Robertson effect: evolutionary
472 consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19–31

473 (2008).
474 24. Pliura, A., Lygis, V., Suchockas, V. & Bartkevicius, E. Performance of twenty four European
475 Fraxinus excelsior populations in three Lithuanian progeny trials with a special emphasis on
476 resistance to Chalara fraxinea. *Baltic For.* **17**, 17–34 (2011).
477 25. Wang, N. *et al.* Genome sequence of dwarf birch (Betula nana) and cross-species RAD
478 markers. *Mol. Ecol.* **22**, 3098–3111 (2013).
479 26. Sollars, E. S. A. *et al.* Genome sequence and genetic diversity of European ash trees.
480 *Nature* **541**, 212–216 (2017).
481 27. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
482 28. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
483 *arXiv [q-bio.GN]* (2013).
484 29. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples.
485 *bioRxiv* 201178 (2018) doi:10.1101/201178.
486 30. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
487 datasets. *Gigascience* **4**, 7 (2015).
488 31. Whittaker, J. C., Thompson, R. & Denham, M. C. Marker-assisted selection using ridge
489 regression. *Genet. Res.* **75**, 249–252 (2000).
490 32. Pérez, P. & de los Campos, G. Genome-Wide Regression and Prediction with the BGLR
491 Statistical Package. *Genetics* **198**, 483–495 (2014).
492 33. Meuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of Total Genetic Value
493 Using Genome-Wide Dense Marker Maps. *Genetics* **157**, 1819–1829 (2001).

494

495 **Acknowledgements**

496

497 Whole genome sequencing of trees in Marden Park wood was funded by a Future Proofing
498 Plant Health grant to R.J.A.B. at RBG Kew from the Department for Environment Food and
499 Rural Affairs (Defra). C.L.M.'s PhD studentship was funded by Defra. Post-doctoral research by
500 C.L.M. was funded by Defra under the Centre for Forest Protection. W.J.P.'s PhD studentship
501 was funded by Defra and Teagasc. We thank the Woodland Trust for access to Marden Park
502 wood, and Tim Hodges and Kate Harvey for their support and assistance.

503

504 This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.
505 http://doi.org/10.5281/zenodo.438045

506

507 **Author contributions**

508

509 R.J.A.B., R.A.N. and C.L.M. conceived the project and designed the analysis. R.J.A.B. selected
510 the field site. R.A.N. and C.L.M. performed the bioinformatic and statistical analyses. C.L.M,
511 W.J.P., J.J.S., L.J.K., M.N.G., R.A.N. and R.J.A.B. conducted the fieldwork. W.J.P. and C.L.M.
512 extracted DNA. J.M. and C.L.M. mapped the trees. L.J.K. advised on analyses. R.J.A.B.,
513 R.A.N., L.J.K.and C.L.M. wrote the paper. R.J.A.B. and R.A.N. supervised the project.

514

515 **Competing interests**

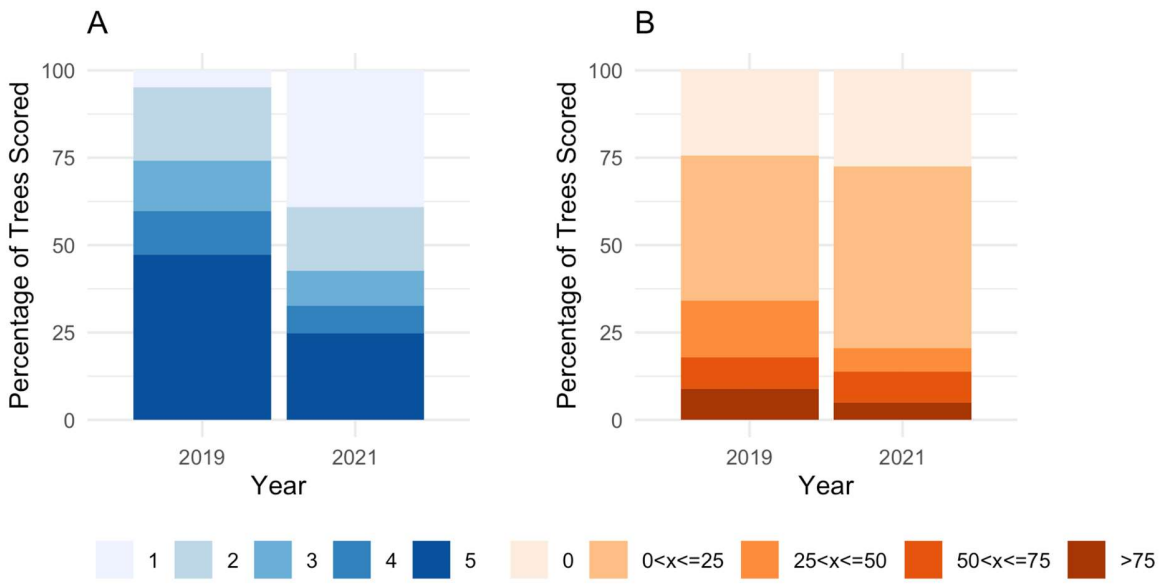
516

517 The authors declare no competing interests.

518

519 **Display Items**

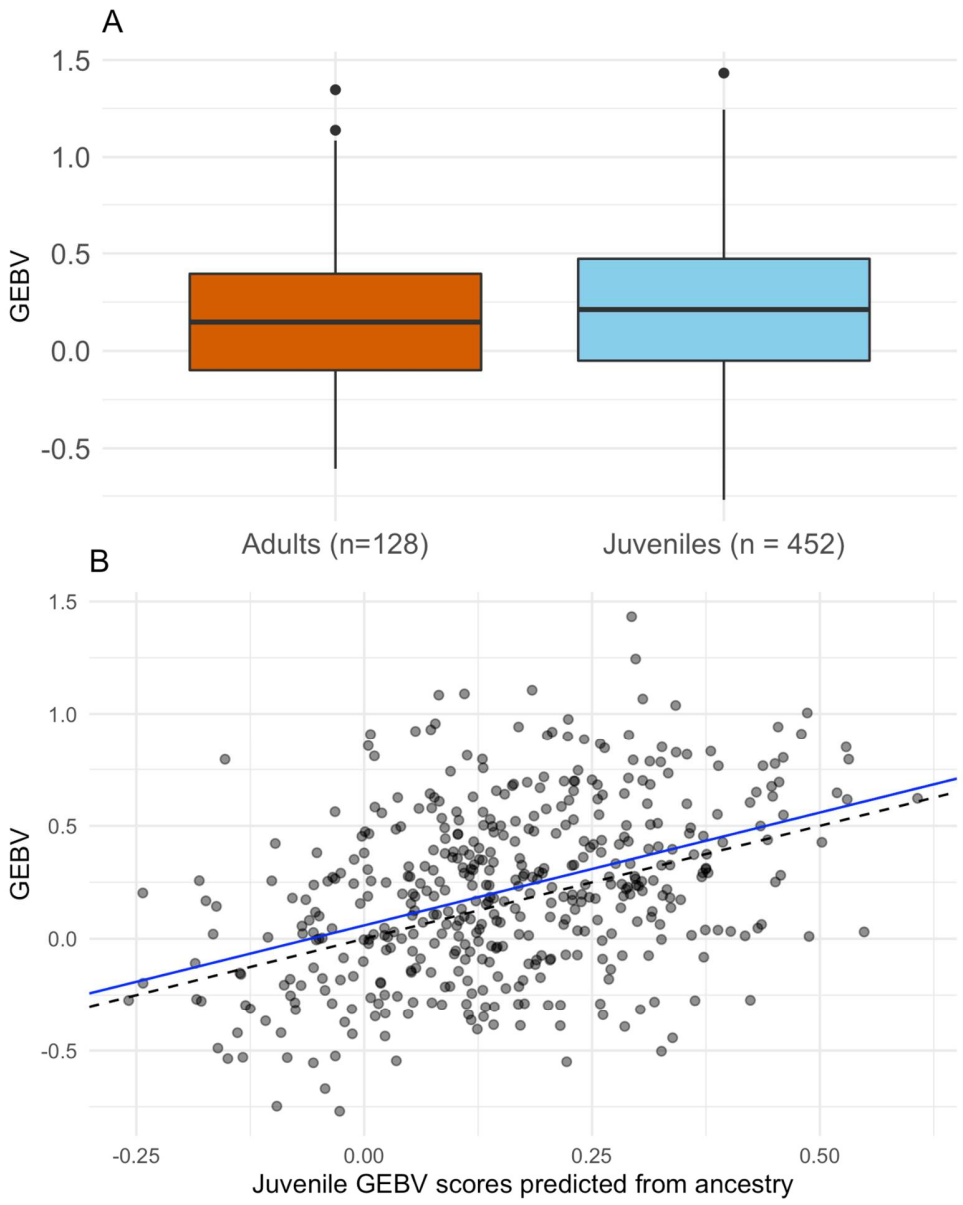
520



521

522 **Figure 1. Phenotypic estimates of ash dieback damage as a percentage of trees scored**
523 **each year.** A) Health scores (where 5 is the most healthy) of juvenile trees as a percentage of
524 the total trees scored for each age group in 2019 and 2021 and B) Percentage canopy coverage
525 estimates of adult trees

526



527

528 **Figure 2. Shift in GEBV between adult and juvenile trees** (A) Genomic estimates of breeding
529 values for 128 adult trees and 452 juvenile trees calculated from 7,985 SNP loci. (B) Juvenile
530 GEBV plotted against the predicted value. The predictions were based on their ancestry using a
531 rrBLUP model trained on the adult trees. The dashed black line indicates the 1:1 line and blue
532 line the fitted regression.

533

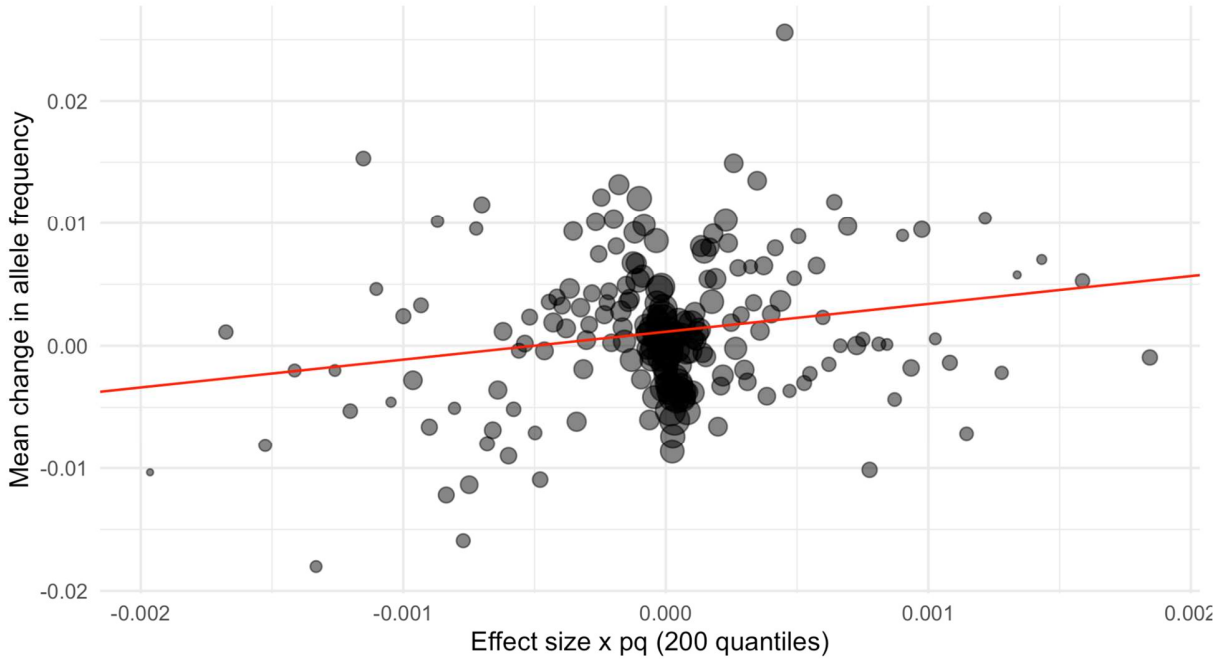


Figure 3. Mean change in allele frequency between adults and juveniles

Mean change in allele frequency plotted against the product of effect size and minor allele frequency. The values for 7,985 sites have been combined into 200 points. Each point represents the average of 39 or 40 sites (adjacent values of Effect size x pq, binned), with area representing the precision of the mean change (1/var(Δf). The red line is the fitted linear regression.

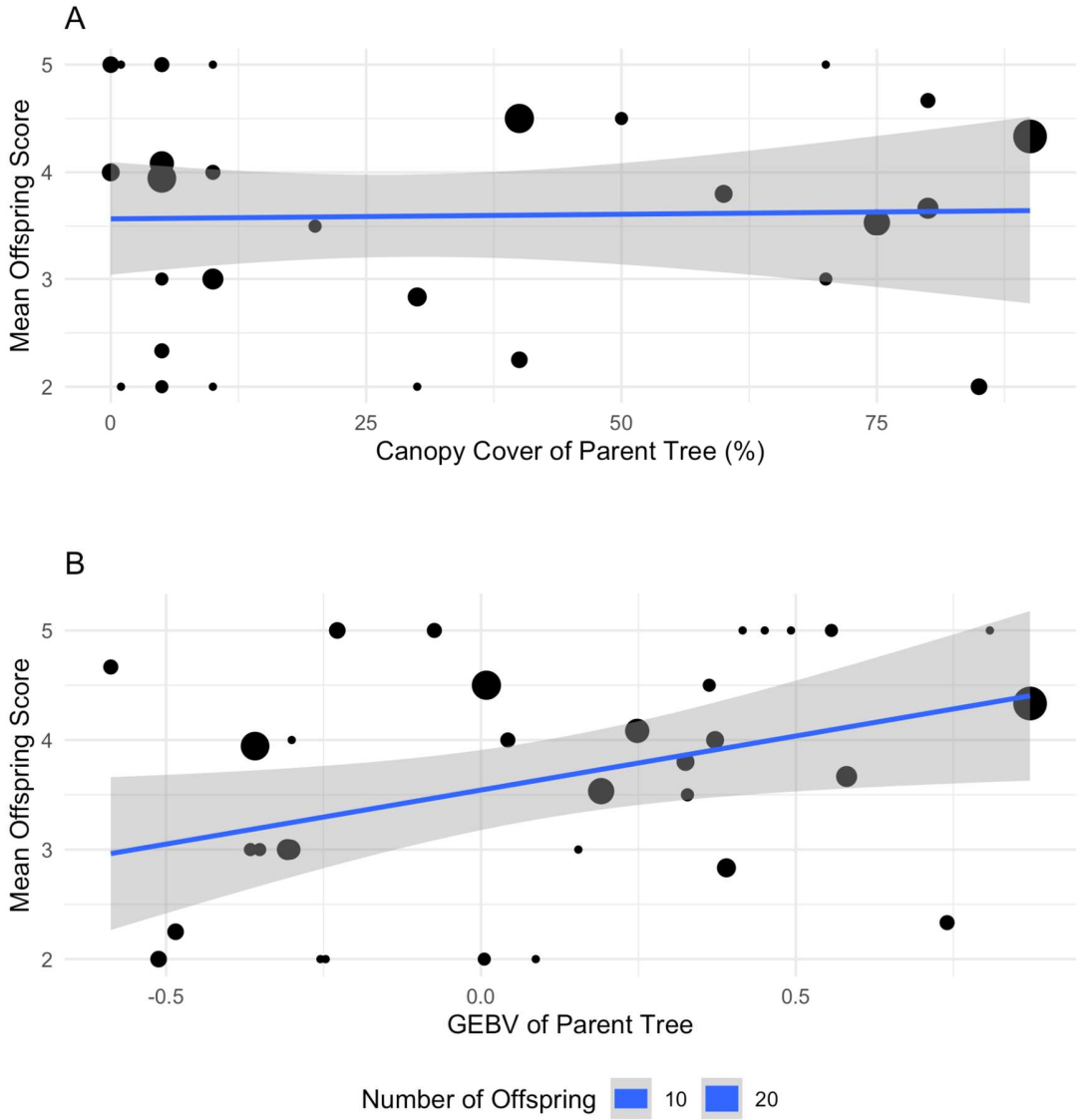
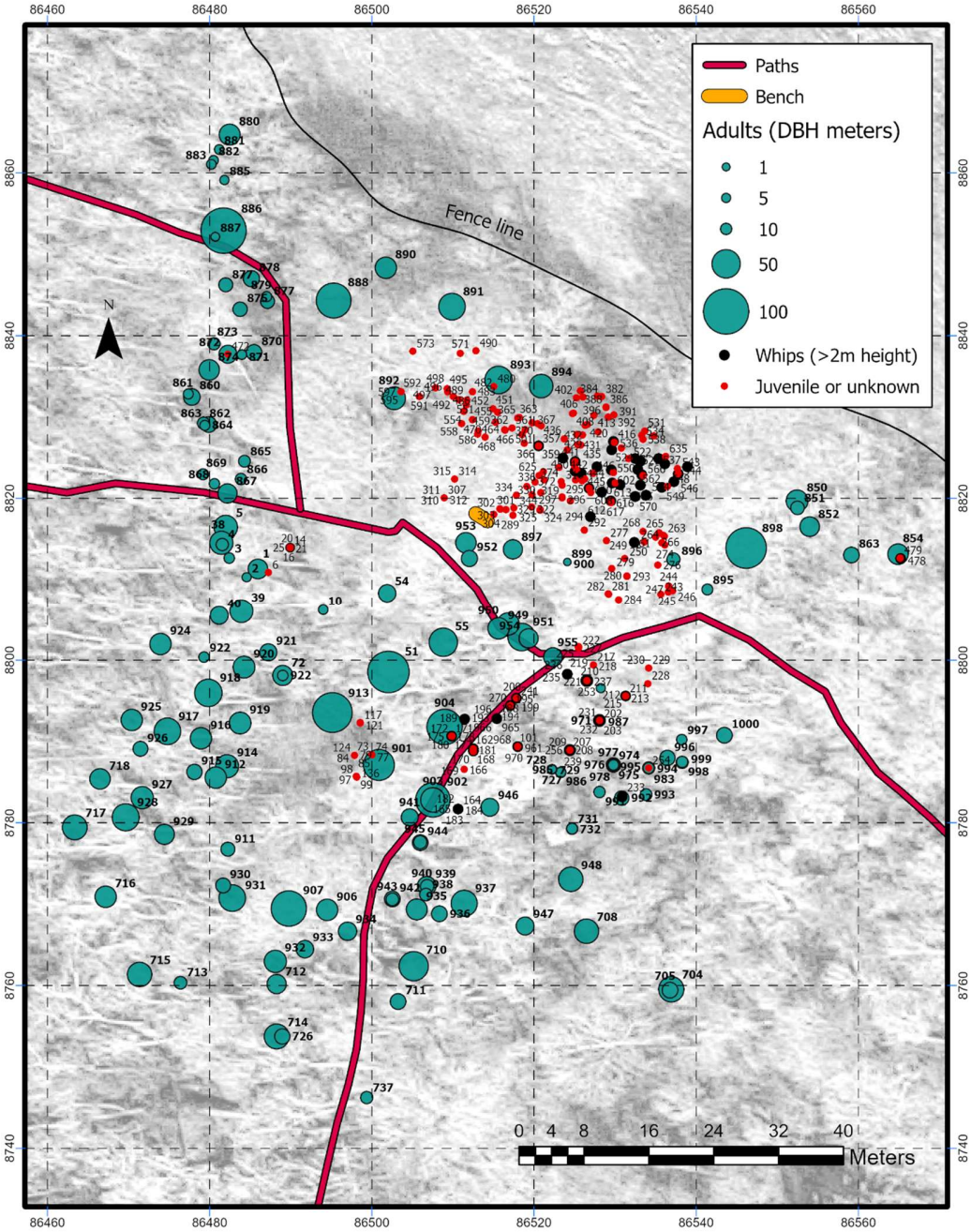
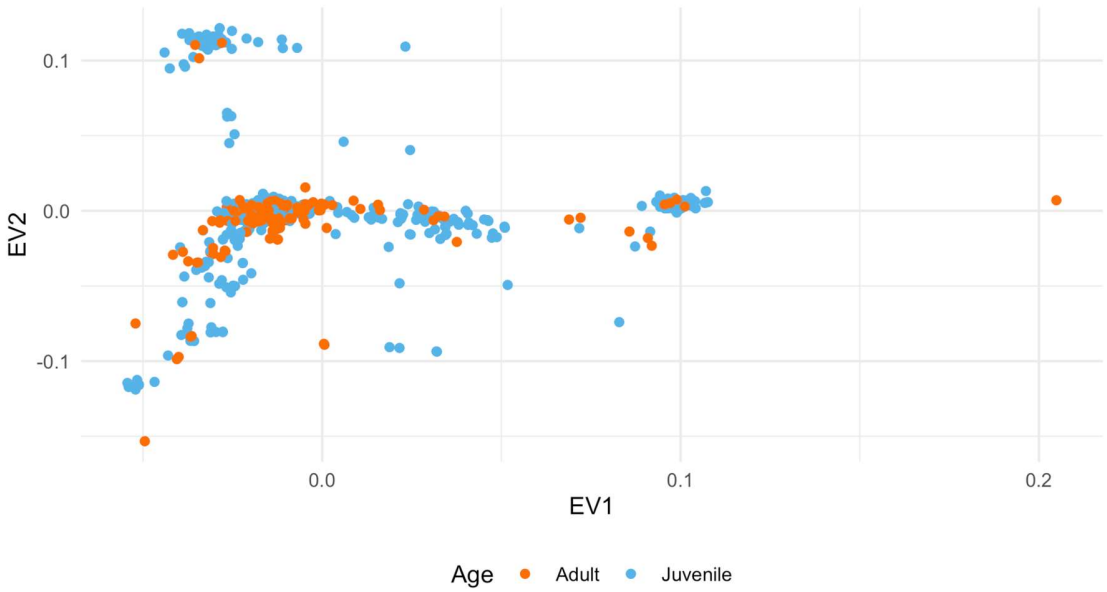


Figure 4. Correlation of A) Canopy cover of likely parent trees and B) GEBV of likely parent trees with the mean phenotypic health score of juvenile offspring as assigned in sequoia. Weighted by the number of offspring trees assigned to each parent.

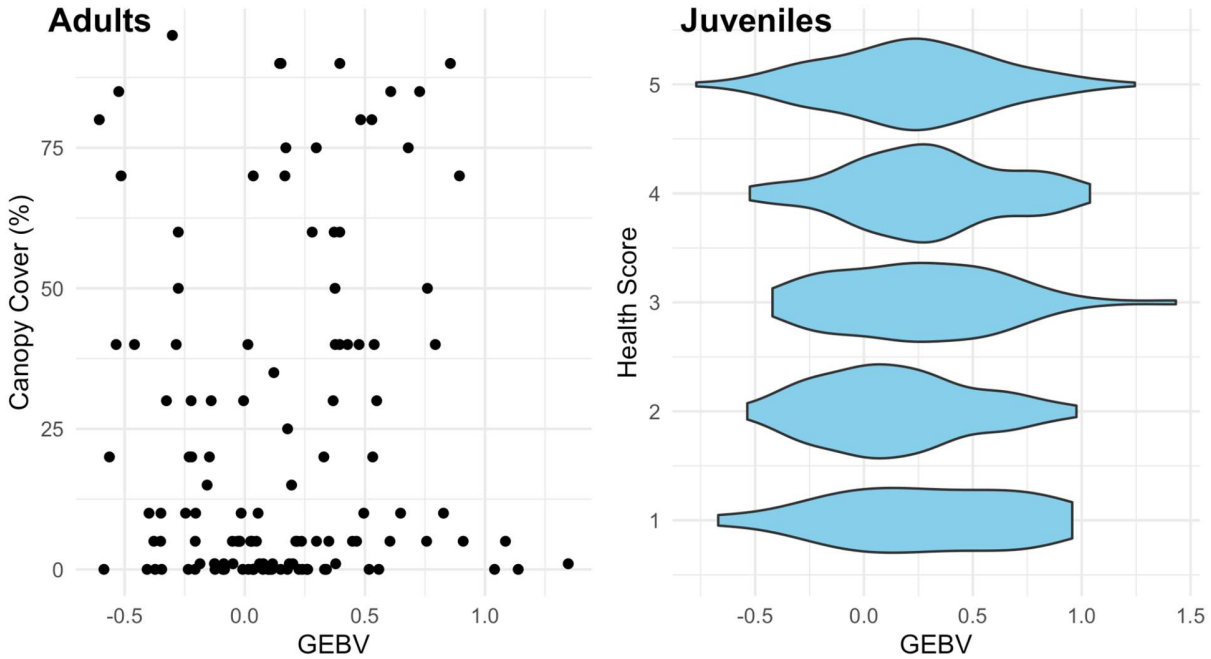
Extended Data



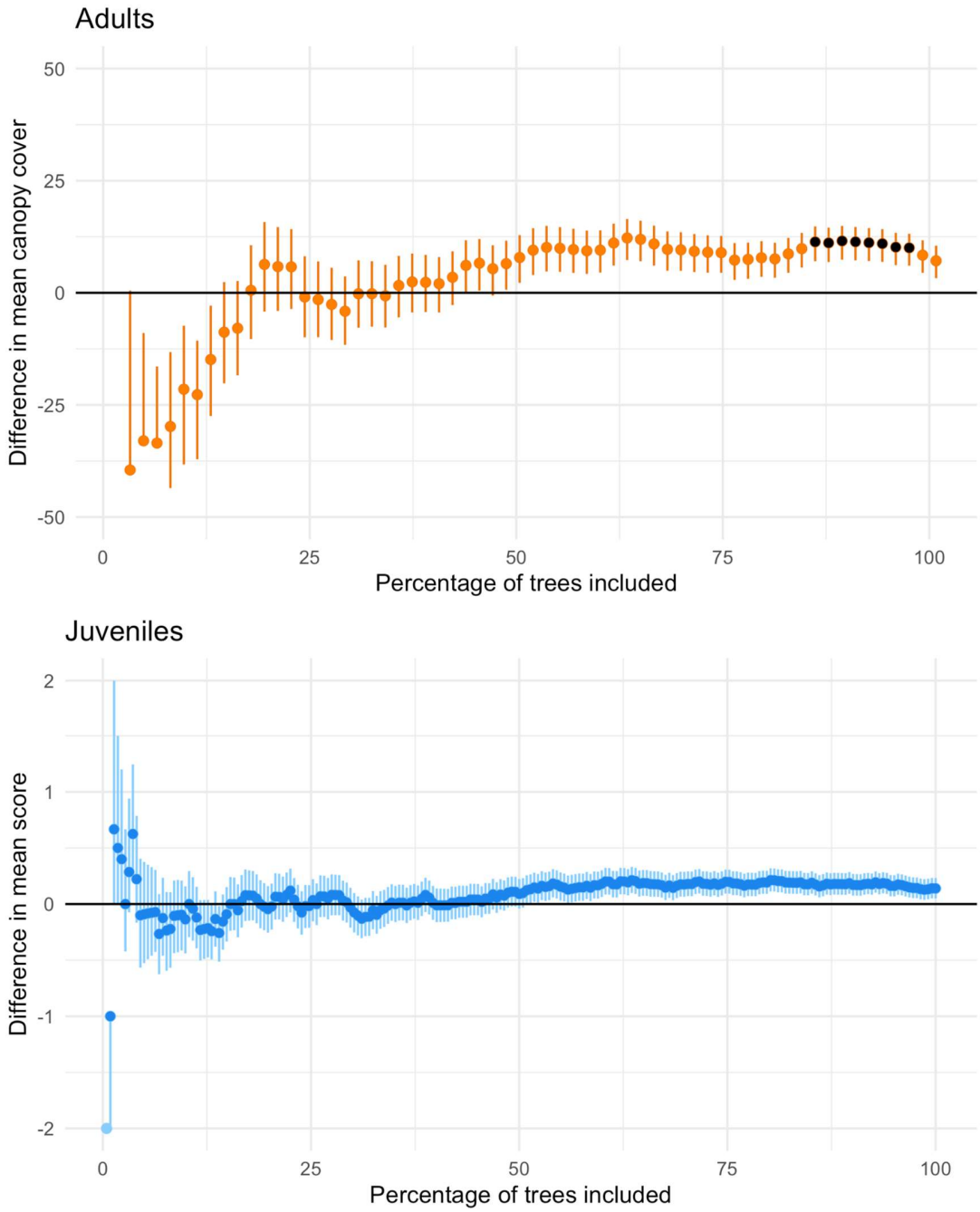
Extended Data Fig. 1. Map of samples showing location of trees at the Marden Park site. Collected in March 2021 using an RTK GPS (Bad Elf Pro), using NTRIP (Networked Transport of RTCM via Internet Protocol) connection to the Imperial University unit (LICCOOGBRO), approximately 27 km from the site.



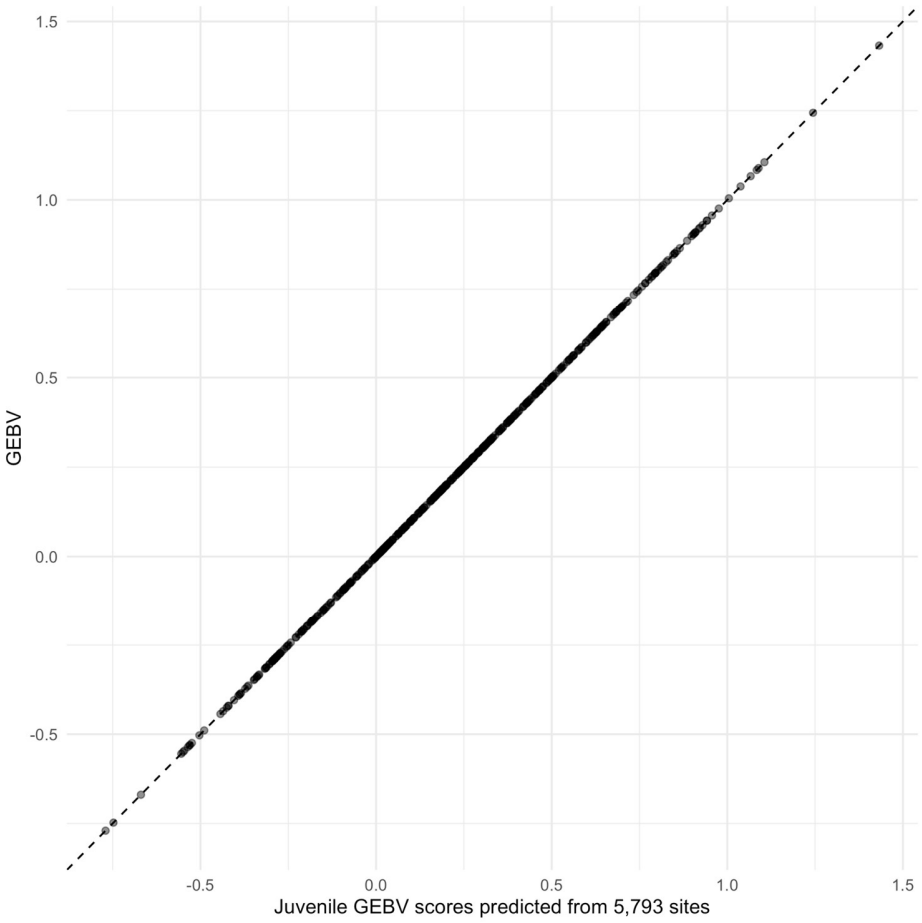
Extended Data Fig. 2. Principal genetic components of all adult and juvenile ash trees, including technical replicates, based on 4,287,210 SNPs.



Extended Data Fig. 3. **Relationship between genomically estimated breeding values (GEBV) and phenotypic estimates of ash dieback damage.** (A) GEBV of 128 adult trees plotted against percentage canopy cover as scored in 2019 and (B) GEBV of 452 juvenile trees plotted against 2019 phenotypic scores on a 1-5 scale where 1 is most severely affected. GEBVs were calculated from 7,985 SNP loci.



Extended Data Fig. 4. **Difference in phenotypic scores of ash dieback damage for trees in equally sized upper and lower extremes of the GEBV distribution.** (A) Percentage canopy cover of adult trees (n = 128) and (B) Health scores of juvenile trees (n = 452). GEBVs were calculated from 7,985 SNP loci. Bars represent the standard error on the cumulative mean phenotypes and points in black show a significant difference ($p < 0.05$) in mean phenotype between the trees in the upper and lower extremes of the GEBV distribution.



Extended Data Fig. 5. Prediction of juvenile GEBV scores using an rrBLUP model trained on 5,793 sites not used in original GEBV calculations produced a perfect correlation (r=1) with the juvenile GEBV scores.

580 **Supplementary information**

581

582 **Supplementary Tables**

583

584 **Supplementary Table 1 | Technical replicate identification.** Genomic distance (1 minus
585 identity-by-state) between replicate samples

Sample 1	Sample 2	Genomic Distance
S31C	S31R	0.067
S54Q	S54R	0.051
S67C	S67R	0.059
S76C	S76R	0.055
S82C	S82R	0.057
S223C	S223R	0.067
S332C	S332Q	0.053
S410C	S410Q	0.061
S415C	S415R	0.059
S571C	S571Q	0.064
S655Q	S655R	0.065
S868C	S868R	0.058
S885C	S885R	0.076
S895M	S895R	0.062
S901C	S901Q	0.067
S925C	S925Q	0.094

S928C	S928Q	0.090
S939C	S939Q	0.061
S954C	S954R	0.078
S987C	S987Q	0.056
A	S612Q	0.059
B	S873Q	0.057
C	121*	
D	5*	
E	S165C	0.063
F	968*	
G	277*	
H	S439R	0.052
I	999*	
J	S202C	0.54

*For half the blind replicates their duplicate did not pass read alignment thresholds and was therefore not included in SNP calling. In these cases the replicate was relabelled to act as a replacement sample.

591 **Supplementary Table 2 | Potential Close Relatives.** Genomic distance (1 minus identity-by-
592 state) between similar samples not labelled as replicates
593

Sample 1	Sample 2	Genomic Distance
7	17	0.074
16	110	0.059
44	951	0.053
156	160	0.061
207	209	0.069
281	282	0.074
281	283	0.073
282	283	0.065
458	461	0.058
470	560	0.069
472	877	0.063
473	476	0.064
533	549	0.064
903	906	0.064

594
595

596 **Supplementary Table 3 | Parentage assignments.** Showing the number and mean (where
597 number of offspring > 1) health scores of offspring assigned to each likely parent tree

Parent Tree	DBH of parent (cm)	% canopy cover in parent (2019)	Parental GEBV	Number of likely offspring	Mean score of offspring (2019)
51	80.0	90	0.8720	29	4.3
10*		5	0.0087	20	4.5
889	42.2	5	-0.3587	19	3.9
39	37.0	75	0.1908	15	3.5
71	49.8	5	0.2481	12	4.1
53	48.1	80	0.5806	8	3.7
890	33.0	10	-0.3074	8	3.0
937	45.0	10	-0.3026	7	2.8
853	39.3	30	0.3898	6	3.0
860	22.3	0	0.3718	5	4.0
945	28.3	60	0.3248	5	3.8
850	34.5	40	-0.4849	4	2.3
898	78.4	85	-0.5119	4	2.0
901	55.5	0	-0.2281	4	5.0
872	24.7	5	0.7402	3	2.3
902	64.0	10	0.0427	3	4.0
907	66.0	80	-0.5879	3	4.7
920	34.0	5	-0.0741	3	5.0

Parent Tree	DBH of parent (cm)	% canopy cover in parent (2019)	Parental GEBV	Number of likely offspring	Mean score of offspring (2019)
51	80.0	90	0.8720	29	4.3
10*		5	0.0087	20	4.5
38	37.5	5	-0.3661	2	3
55	50.1	50	0.3625	2	4.5
884	89.8	70	-0.3513	2	3
886	40.0	5	0.0052	2	2
908	4.5	0	0.5565	2	5
910	4.5	20	0.3279	2	3.5
52	35.0	5	-0.2242	1	NA
712	28.0	NA	0.4506	1	5
852	29.6	10	-0.2466	1	2
896	14.2	1	0.0870	1	2
903	40	0	0.2233	1	NA
927	35.0	1	0.4155	1	5
928	47.0	30	-0.2549	1	2
934	26.0	5	0.1546	1	3
973	6.8	10	0.4924	1	5
978	13.2	0	-0.3005	1	4
988	16.8	0	-0.0603	1	NA

Parent Tree	DBH of parent (cm)	% canopy cover in parent (2019)	Parental GEBV	Number of likely offspring	Mean score of offspring (2019)
51	80.0	90	0.8720	29	4.3
10*		5	0.0087	20	4.5
989	4.1	70	0.8080	1	5

* Accurate DBH measurement missing for tree 10

Supplementary Data 1. Phenotypic data for all trees scored or sampled at Marden Park

Supplementary Data 2. Effect sizes of 7,985 sites used to calculate GEBV

Supplementary Data 3. Genotype frequency matrix for 7,985 sites used to calculate GEBV