Letter

Amborella gene presence/absence variation is associated with abiotic stress responses that may contribute to environmental adaptation

Amborella trichopoda (Amborellaceae) is the single living sister species of all other extant flowering plants and only occurs in rain forest habitats on the remote island of New Caledonia. These features make *Amborella* an important species in which to study genetic variation, including gene presence/absence variants (PAVs). Here, we apply the reference genome based iterative mapping and assembly strategy (Bayer *et al.*, 2020) to assess gene diversity across 10 diverse individuals. The N50 of the newly assembled contigs was 1.2 kb, indicating similar contiguity to those in comparable pan-genome studies (*Brassica oleracea*, 0.6–1.9 kb; *Brassica napus*, 1.2 kb; *Solanum lycopersicum* L., 1.4 kb; *Musaceae*, 0.7–2.4 kb; and *Oryza sativa* L., 1.1 kb) (Golicz *et al.*, 2016a; Hurgobin *et al.*, 2018; Wang *et al.*, 2018; Gao *et al.*, 2019; Rijzaani *et al.*, 2021) (Supporting Information Table S1). We identified 2765 additional genes not present in the reference assembly and found that *Amborella* may have relatively few dispensable genes (3136, 10.4%) (Table S2) compared with studies in other species (maize, 60.88%; *Brassica oleracea*, 18.71%; bread wheat, 42.30%; *Brachypodium distachyon*, 45%; and cultivated rice, 51.5%) (Hirsch *et al.*, 2014; Golicz *et al.*, 2016a; Gordon *et al.*, 2017; Montenegro *et al.*, 2017; Wang *et al.*, 2018). Although a small set of samples was included in this study, our pan-genome modelling indicates that 11 *Amborella* samples were sufficient to capture the majority of PAVs within *Amborella* (Fig. S1). Although levels of genetic diversity in *Amborella* are comparable with outbreeding perennials such as *Populus*, population bottlenecks over the past 900 000 yr may have contributed to the relatively low number of gene PAVs (Amborella Genome Project, 2013).

Studies in *Brassica oleracea* (e.g. cabbage, broccoli, cauliflower, kale), wheat, rice, tomato, soybean and maize have demonstrated that dispensable genes are frequently associated with both biotic and abiotic stress (Golicz *et al.*, 2016b; Jin *et al.*, 2016; Montenegro *et al.*, 2017; Wang *et al.*, 2018; Gao *et al.*, 2019; Liu *et al.*, 2020). By contrast, the dispensable genes in the *Amborella* genome are mainly enriched for functions associated with responses to abiotic stress, including salt, cadmium, zinc, cold, water deprivation and heat, with few dispensable genes associated with biotic stress (Tables S3,S4). Based on Gene Ontology (GO) annotation, 314 genes are related to cadmium ion response in *Amborella*, which is similar to 296 cadmium ion response genes identified in *Brassica oleracea* and significantly higher than the number of cadmium response genes identified in *Brachypodium distachyon* (44), *Oryza sativa* L. (44), *Solanum lycopersicum* L. (60) and *Glycine max* (28) (Tables S5,S6). *Amborella* has the highest proportion of cadmium-responsive genes classified as dispensable (44.6%), compared with *Brassica oleracea* (7.8%), *Glycine max* (7.8%), *Brachypodium distachyon* (25.0%), *Oryza sativa* L. (22.7%) and *Solanum lycopersicum* L. (23.3%) (Table S6).

One-third of New Caledonia is covered by ultramafic soils that are rich in nickel, lead, silver, zinc, copper and cadmium (Lillie & Brothers, 1970). Although *Amborella* does not occur on ultramafic soils, it does occur on soils rich in metals (Thien *et al.*, 2003), and genes dispensable in *Amborella* may be associated with adaptation to regions of the island with high metal content soils (Jaffre *et al.*, 2013). Cadmium ion response genes can be grouped based on Pfam domain classification into 97 gene families, with 57 families only containing one single gene and 40 gene families containing multiple gene copies (Fig. S2; Table S7). Different gene families show expansion or contraction across the *Amborella* samples. For example, 12 cadmium ion response-associated genes have a Myb-like DNA-binding domain (PF00249); only seven copies were identified in the reference sample Santa Cruz, while all 12 copies were found in samples BA, BO, TOB and PWB (Fig. S3). We assessed more broadly the distribution of cadmium ion response-associated genes, showing that many are missing in the reference Santa Cruz, Mé Ori (MO) and PO samples (Fig. S4). However, the samples TC, BO, AM, TOB, BA and PWB experienced significant gene expansion. These samples were located near two ancient refugia (Poncet *et al.*, 2013), suggesting that the observed difference in gene content may be associated with population changes during the last glacial maximum (*c.* 21 000 yr BP).

We identified 152 dispensable genes predicted to be involved in the salt response, 100 related to the cold response, and 76 associated with the drought response (Fig. S2; Table S5). The geographical features and climate conditions vary significantly across New Caledonia. A central mountain range results in a rain shadow effect, with 800 mm yr^{-1} in the western coastal region compared with 4500 mm yr^{-1} on the eastern slopes of mountain areas (Teurlai *et al.*, 2015). Moreover, as the average annual temperature decreases by *c.* 0.6°C with every 100 m increase of elevation (Andréfouët *et al.*, 2004), *Amborella* populations occurring in elevations ranging from 110 m to 860 m (Poncet *et al.*, 2013) grow under different temperature conditions. The diverse environmental conditions present within a relatively small and isolated island may be a driver for the retention or loss of genes for environmental adaptation across the island (Hodel *et al.*, 2018).

To also assess the diversity of resistance gene analogue (RGA) content in *Amborella*, we conducted genome-wide RGA discovery analysis. Over-masking repeats in a genome assembly can lead to a

decreased number of RGAs identified (Bayer *et al.*, 2018). Even without masking repeats, we found that *Amborella* contained relatively few RGAs (514), compared with other angiosperms of similar genome size (*Solanum lycopersicum* L., 1000; *Manihot esculenta*, 1412; *Brassica napus*, 1749; and *Brassica oleracea*, 1989) (Li *et al.*, 2016; Bayer *et al.*, 2019; Dolatabadian *et al.*, 2020). Most RGAs were receptor-like kinase genes (RLK; 308 genes), followed by nucleotide binding site leucine-rich repeat genes (NLR; 140 genes) and receptor-like proteins (RLP; 68 genes) (Fig. S5; Table S8), consistent with previous reports that RLKs are the most abundant class of RGAs in plants (Shiu *et al.*, 2004). The physical clustering of RGAs (Table S9) is also similar to observations in other species (Bayer *et al.*, 2019).

NLRs, which play an important role in disease resistance responses, are divided into subclasses based on their domain structures. Toll/interleukin-1-Nucleotide-binding site-Leucine-rich repeat (TNL) and Coiled-coil-Nucleotide-binding site-Leucine-rich repeat (CNL) are the two typical complete forms of NLRs. The remaining NLR subclasses, with missing domains and disordered domain structure, are known as atypical NLRs. *Amborella* contained 28 typical NLRs (CNL, 14; TNL, 14) and 112 atypical NLRs, of which the majority were NL (40) and NBS (39). Zhang *et al.* (2016) identified a total of 88 NLRs in *Amborella*, including 9 TNLs and 15 CNLs. However, Shao *et al.* (2016) reported a larger number of CNLs (89) and TNLs (15) in *Amborella*. The differences in the numbers of RGAs reported in these studies were due to the application of different RGA classification approaches (Tables S10, S11). Specifically, Shao *et al.* (2016) defined all the non-TNL RGAs as CNL genes, while we include the subclasses CC-NBS-LRR (CNL), CC-NBS (CN), NBS-LRR (NL), and NBS (NBS). In addition, our RGA analysis relies on the *A. trichopoda* v.6.0 assembly and gene annotation as well as newly annotated pan-genome genes, whereas Shao *et al.* (2016) and Zhang *et al.* (2016) rely on the lower number of genes in the *A. trichopoda* v.1.0 gene annotation. Overall, our analysis captures 105 of the 108 previously identified RGAs as well as detecting 35 previously unreported RGAs (Fig. S6), including 4 TNLs (Table S12). Of the 514 RGAs, we found that most RGAs (491) were core and the remaining 23 were dispensable. This result contrasts with findings in other studies in which the majority of RGAs are dispensable (Li *et al.*, 2014; Hurgobin *et al.*, 2018; Bayer *et al.*, 2019). The relatively low number and lack of dispensable RGAs in *Amborella* may, in part, reflect the lack of diverse pathogen pressure given its isolated and limited distribution on a remote island, as well as the lack of recent WGD events, which tend to increase both the number and variation of NLRs across a genome (Seo *et al.*, 2016).

Lastly, we explored the population structure of *Amborella* using SNP-based and gene PAV-based population analysis. The 10 *Amborella* individuals in this study were sampled from across the native range in New Caledonia (Fig. 1), while the reference sample Santa Cruz was sourced from the California Santa Cruz Arboretum and Botanic Garden, with evidence suggesting that it had a similar origin as MO (Amborella Genome Project, 2013). A SNP-based phylogenetic tree and principal component analysis (PCA) plot (Fig. 2a,b) showed that *Amborella* individuals clustered into four

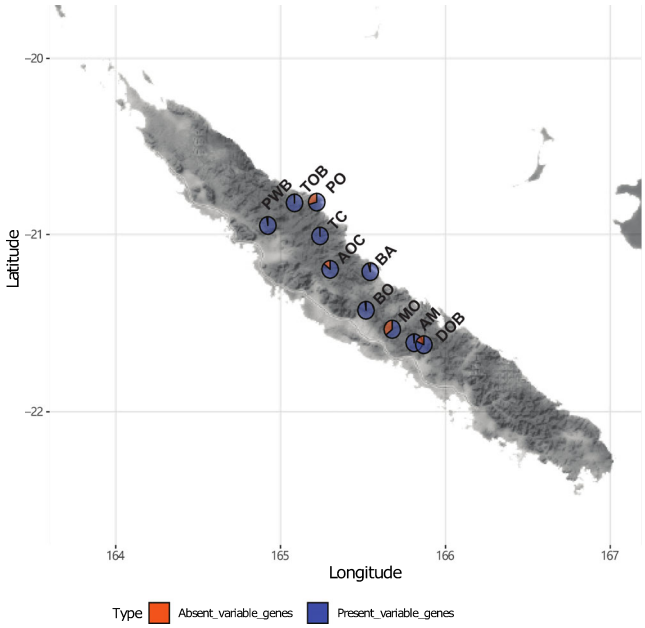

Fig. 1 Percentage of presence/absence of dispensable genes in 10 different *Amborella* individuals growing in different geographical locations in New Caledonia. The following abbreviations are used to describe the geographical locations. AM, Amieu; AOC, Aoupinié; BA, Ba Houailou; BO, Boregaou; DOB, Dogny; MO, Mé Ori; PO, Ponandou; PWB, Pwicate; TC, Tchamba; TOB, Tonine.

clades or groups based on their geographical locations, consistent with previous studies that inferred *Amborella* population structure using microsatellite loci (Poncet *et al.*, 2013) and SNPs (Amborella Genome Project, 2013). However, cluster and PCA analyses based on PAVs showed some differences in population structure (Fig. 2c,d). The two individuals Santa Cruz and MO remained closely associated, although the PCA based on PAVs suggested greater variation between these individuals than found by the SNP analyses. Ponandou (PO), with the second fewest genes, is located on the north-eastern side of the island and showed a distinct PAV pattern compared with Santa Cruz and MO (Fig. 2d). The cluster containing DOB, BA and AOC showed an intermediate pattern, while the third cluster (comprising BO, AM, TC, TOB and PWB) hosted the largest number of newly annotated genes (Figs 1, S7; Table S13).

The differences in gene content between clusters may be driven by environmental factors and past ecological processes. The individuals in the cluster comprising BO, AM, TC, TOB, PWB, DOB, BA and AOC shared the majority of the newly annotated genes. These samples were collected from the northern and central areas of New Caledonia, two putative refugial areas from the last glacial maximum with higher levels of genetic diversity than other areas (Poncet *et al.*, 2013). Although the MO sample is geographically close to AM and DOB, it shares a low number of genes with the reference sample Santa Cruz; this differentiates MO from other samples. This gene loss and low diversity may have been caused by either a bottleneck or a founder effect at the fringe of the geographical distribution (Hampe & Petit, 2005). Additionally, PAVs and SNPs are suitable for inferring population structure because they often show identity by descent as a result of single

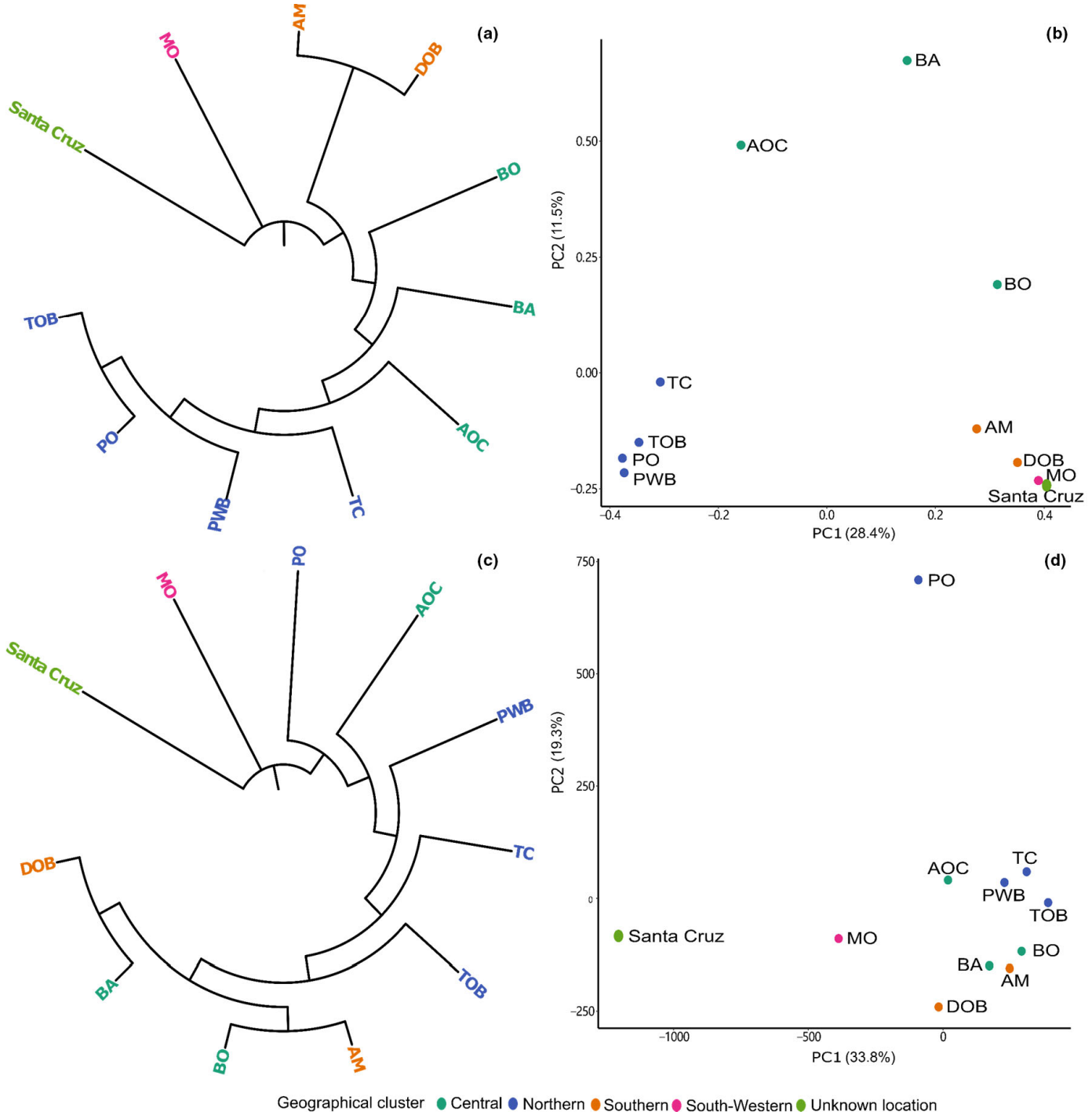

Fig. 2 Maximum likelihood phylogenies and PCA plots showing the relationship between *Amborella* individuals. (a) SNP-based phylogeny; (b) SNP-based PCA plot; (c) PAV-based phylogeny; (d) PAV-based PCA plot. Colours in phylogenetic trees and PCA plots are based on the geographical location cluster. The following abbreviations are used to describe the geographical locations. AM, Amieu; AOC, Aoupinié; BA, Ba Houailou; BO, Boregaou; DOB, Dogny; MO, Mé Ori; PO, Ponandou; PWB, Pwicate; TC, Tchamba; TOB, Tonine.

ancestral mutation events (Sudmant *et al.*, 2015). However, PAVs may lead to different outcomes when used in phylogenetic inference compared with SNPs. For example, in *Arabidopsis*, phylogenies based on SNPs and PAVs found congruent relationships between major geographic clades but differed substantially in how they resolved relationships within these clades (Tan *et al.*, 2012). Such patterns of PAVs conflicting with evolutionary distance and geography may also be caused by balancing selection and highlight the value of using PAVs in addition to SNPs in assessing species evolution.

By characterising genetic diversity in the phylogenetically pivotal plant *Amborella*, our study showed that dispensable genes are

annotated with functions associated with abiotic stress response that have the potential to contribute to adaptation. The 12 known Amborella populations, 10 of which we investigated here, occur in distinct sites distributed across New Caledonia (Amborella Genome Project, 2013; Poncet *et al.*, 2013), and their genetic structure reflects their geographical location. Although some data on environmental conditions are available, the small number of samples hinders robust statistical inference of associations between gene PAVs and the environment. Further sampling of within-population diversity in *Amborella*, together with a more detailed assessment of the environment, may help to elucidate the effect of environment on genomic variation. However, Poncet *et al.* (2013) reported high genetic homogeneity within *Amborella* populations, suggesting that additional sampling would not uncover significantly more diversity. Even with limited sampling, our gene presence/absence analysis identified variation in *Amborella* gene family size that may be associated with local adaptation to the environment. By contrast with many other plants, particularly crops, *Amborella* resistance gene family sizes are relatively small and invariable. Notably, however, there was a substantial variation in cadmium-responsive genes and other genes responsive to environmental stress. This suggests that structural variation has the potential to contribute to adaptation in *Amborella* and families of dispensable environmentally responsive genes may be important for such adaptations.

Materials and Methods

Sequencing

Pennsylvania State University provided whole-genome Illumina re-sequencing data for *Amborella trichopoda* as part of the *Amborella* Genome Project. One *Amborella trichopoda* individual from each of 10 natural populations represented the geographic and genetic diversity of *Amborella* on the mainland of the New Caledonia archipelago. These sites were sampled based on Poncet and colleagues' microsatellite analysis (Poncet *et al.*, 2013). In addition, sequence data for one plant from the University of California Santa Cruz (Santa Cruz) were included in the reference genome (Amborella Genome Project, 2013). The data published by the *Amborella* Genome Project are available in the NCBI Sequence Read Archive (SRA), accessions SRX337107–SRX337118 and SRX336900. The geographic locations are presented in Table S14.

Novel contig assembly

Here, 10 individuals with sequencing coverage of >10× were selected for novel contig assembly using a previously published assembly pipeline (Golicz *et al.*, 2016a). An updated chromosome-level *Amborella* assembly was used as the starting reference. Iterating through each of the 10 samples, reads were aligned to the reference genome and unaligned reads were assembled into contigs. These contigs were then added to the reference before the next sample was aligned. The *Amborella* chloroplast genome (NC_005086.1) and mitochondrial genome (KF754803.1) were included in the reference assembly. Read adapters were removed using Trimmomatic (Bolger *et al.*, 2014) v.0.36. Alignment was performed using Bowtie2 (Langmead & Salzberg, 2012) v.2.3.3.1 with parameters (--end-to-end --sensitive -I 0 -X 1000), and the *de novo* assembly of the unaligned reads was conducted using MaSuRCa (Zimin *et al.*, 2013) v.3.2.2. The resulting pan-genome was assessed using reads re-aligned by Bowtie2.

Removal of contaminants

Blast+ (Camacho *et al.*, 2009) v.2.2.31 (-template_type coding_and_optimal -max_target_seqs 2 -*e*-value 1e-3) was used to perform contamination detection, by comparing the novel contigs with the NCBI nucleotide database (downloaded from NCBI, 26 June 2018). Query sequences showing best hits with over 80% query length covered and 80% identity to likely contaminants (sequences of nongreen plant species) were removed.

Novel contig annotation

To reduce variation in annotation caused by different annotation tools, we used the conservative annotation approach of Evidence-modeler (EVM) (Haas *et al.*, 2008) v.1.1.1 to annotate the novel contigs, following the approach of the *Amborella* reference genome project (Amborella Genome Project, 2013). Three sets of evidence (gene prediction evidence, transcript evidence and protein evidence) were provided as the inputs. Gene evidence of newly assembled contigs was generated by running the Snap (Korf, 2004), Augustus (Stanke *et al.*, 2006) and Genemark (Tang *et al.*, 2015) gene prediction pipelines. For the transcript evidence, the published available RNA-seq data from reference sample were downloaded from NCBI (Table S15). Tophat (Trapnell *et al.*, 2009) v.2.1.1 and Cufflinks (Roberts *et al.*, 2011) v.2.2.1 were used to perform the RNA alignment and assembly. In addition, the published RNA-seq assembly (available from the *Amborella* Genome Project (Amborella Genome Project, 2013)) was aligned using the Pasa transcript prediction pipeline. Protein evidence was generated by aligning the protein sequences of *Amborella* (downloaded from NCBI GenBank) using Exonerate (Slater & Birney, 2005) v.2.2. Moreover, repeat sequences masked by Repeatmasker (Tarailo-Graovac & Chen, 2009) v.4.0.7 were provided as an additional input to assist annotation.

Gene presence/absence variation

Eleven *Amborella* samples (10 from natural populations and one reference sample) were used in gene PAV calling. Gene PAV discovery was based on coverage analysis using the Mosdepth package (Pedersen & Quinlan, 2018). Reads from all lines were mapped to the genome and novel contigs using Bwa-mem with default parameters (Li & Durbin, 2009). Gene PAVs were determined based on the depth-of-coverage calculation across all exons of the genes by Mosdepth. We do not detect significant correlation of the genes present in each line with sequencing depth (Pearson's correlation = 0.27; *P*-value = 0.42) (Table S16). A gene was considered as missing when the horizontal coverage across exons of the gene was <5% and the vertical coverage was <2×

(Golicz *et al.*, 2015). A gene was considered a core gene if it was present in all 11 samples; alternatively, if it was missing in one sample, it was considered a dispensable gene. The R package UPSETR (Conway *et al.*, 2017) was used to show the distribution of gene PAVs among the 11 *Amborella* samples. The change of pan-genome size and core genome size were simulated using the nlsLM function from the package MINPACK.LM (Elzhov *et al.*, 2016).

GO annotation

The pan-genome and protein sequences of rice (*Oryza sativa* L.) (Wang *et al.*, 2018) were downloaded from the Rice Pan-genome Browser (http://cgm.sjtu.edu.cn/3kricedb/index.php). The pan-genome and protein sequences of *Brachypodium distachyon* (Gordon *et al.*, 2017) were downloaded from the JGI *Brachypodium* Pan-genome resource website (https://brachypan.jgi.doe.gov/). The tomato (*S. lycopersicum* L.) pan-genome (Gao *et al.*, 2019) and its protein sequences were downloaded from the Dryad Digital Repository (10.5061/dryad.m463f7k). The *Brassica oleracea* pan-genome (Golicz *et al.*, 2016a) and its protein sequences were downloaded from the *Brassica* genome database (http://brassicagenome.net/databases.php). The cultivated soybean pan-genome (Torkamaneh *et al.*, 2021) and its protein sequences were downloaded from Soybase (https://www.soybase.org/projects/SoyBase.C2021.01.php).

Functional annotation was performed using command line BLAST2GO (Conesa & Gotz, 2008) v.2.5. The genes were aligned to the proteins in the Viridiplantae database using BLASTP (Camacho *et al.*, 2009), and only alignments with *E*-values $< 1 \times 10^{-5}$ were used. Then, the BLAST results were reformatted to satisfy BLAST2GO naming requirements. BLAST2GO was further used to identify GO for aligned genes based on the BLAST results. GO enrichment analysis of the dispensable genes was conducted by the R package TOPGO (Alexa & Rahnenfuhrer, 2010) using Fisher's exact test with the approach 'elim' used to correct for multiple comparisons. Genes assigned with the GO term GO:0046686 (response to cadmium ion) were extracted and used to determine the percentage of the dispensable genes in *Amborella trichopoda*, *Brassica oleracea*, *Glycine max*, *Brachypodium distachyon*, *Oryza sativa* L. and *S. lycopersicum* L.

RGA candidate gene discovery

The RGAUGURY pipeline (v.2017-10-21) (Li *et al.*, 2016) was used to predict NBS, RLK and RLP candidate genes. Resistance gene physical clusters were detected by comparing the genetic location of RGA candidates using the method of Bayer *et al.* (2019). If two RGAs were located within each other by 10 upstream or 10 downstream genes, they were merged into an RGA-gene-rich cluster. For comparison of RGA sets from Shao *et al.* (2016) and Zhang *et al.* (2016) using the *A. trichopoda* v.1.0 gene annotation with our RGA set based on the *A. trichopoda* v.6.0 assembly and pan-genome annotation, we used BLASTP (Camacho *et al.*, 2009) v.2.2.31 to identify orthologues with a sequence similarity threshold of 100%. The RGAUGURY pipeline was used to detect RGA candidates in these orthologues in *A. trichopoda* v.1.0 gene annotation (Amborella Genome Project, 2013). We detected 140 RGAs, including 100% of the 105 RGAs found by Shao *et al.* (2016) and 96.6% of the 85 RGAs found by Zhang *et al.* (2016). We assessed the six genes with a conflict in classification between our results and those in Shao *et al.* (2016), using the coiled-coil prediction software COILS (https://embnet.vital-it.ch/software/COILS_form.html) and Pfam protein domain search (http://pfam.xfam.org/search/sequence) (Table S17). Based on functional protein domains, we found that our classifications were supported (Figs S8–S10).

SNP discovery

Reads were aligned using BWA-MEM (Li & Durbin, 2009) with default settings. The mapped reads were then sorted and duplicates removed by PICARD tools (McKenna *et al.*, 2010). The reads were re-aligned using the GATK REALIGNER TARGETCREATOR and INDELREALIGNER package, followed by variant calling using GATK HAPLOTYPECALLER (McKenna *et al.*, 2010). The resulting SNP variants were filtered (QD < 2.0 || MQ < 40.0 || FS > 60.0 || QUAL < 60.0 || MQrankSum < −12.5 || ReadPosRankSum < −8.0) to remove low-quality SNPs. All the SNP-based analyses were carried out using the filtered SNPs. For calculation of gene SNP densities, there may be a bias towards lower SNP densities in dispensable genes. The dispensable genes, unlike core genes, are by definition not present in all samples. This may lead to a lower number of SNPs discovered that are not due to a true biological effect, but due to a smaller number of lines contributing the SNPs. To perform a more reliable comparison of SNP densities between core and dispensable genes, the SNP density was therefore normalised for the number of gene copies present according to the following formula:

$$\text{Number of SNPs}/(\text{Number of gene copies present} - 1) \times \text{gene length}/10.$$

SNPs were annotated using SNPEFF (Cingolani *et al.*, 2012).

Phylogenetic analysis

A SNP-based phylogenetic tree of 11 *Amborella* accessions was constructed using IQ-TREE (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017) using a maximum likelihood method (with the parameters -alrt 1000 -bb 1000). The PMB + F + I substitution model was selected using MODELFINDER based on the Bayesian Information Criterion. before tree construction, high-confidence bi-allelic SNPs were generated by removing SNPs with minor allele frequency (MAF) < 0.05 and missing genotype rate < 10% using VCFTOOLS (Danecek *et al.*, 2011). The PAV-based maximum likelihood phylogenetic tree was generated with IQ-TREE (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017) (with the parameters -alrt 1000 -bb 1000) using the PAVs binary matrix in phylip format as input (gene presence was recorded as 1 and absence was recorded as 0). The SYM+ASC+R2 substitution model was selected using MODELFINDER based on the Bayesian Information Criterion.

Acknowledgements

We thank members of the *Amborella* Genome Project for allowing us to use the updated chromosome-level assembly in this analysis. This work was undertaken with the assistance of resources provided at the Pawsey Supercomputing Centre and was partially funded by the Australia Research Council (Projects FT130100604, LP160100030, LP140100537 and LP130100925) and the US National Science Foundation (grant no. IOS-0922742). HH thanks the China Scholarship Council for supporting his studies at the University of Western Australia. AS was supported by an IPRS awarded by the Australian government. PB acknowledges the support of the Forrest Research Foundation. The authors declare no competing interests.

Author contributions

DE, PB, JB, DS and PS designed the experiments and coordinated the project. RGJH contributed to the genetic analysis. ST contributed to the disease resistance analysis. BV and AS contributed to the pan-genome assembly analysis. HH performed all other computational analyses. All authors contributed extensively to the writing and editing of the manuscript.

ORCID

Jacqueline Batley https://orcid.org/0000-0002-5391-5824
Philipp E. Bayer https://orcid.org/0000-0001-8530-3067
David Edwards https://orcid.org/0000-0001-7599-6760
Richard G. J. Hodel https://orcid.org/0000-0002-2896-4907
Haifei Hu https://orcid.org/0000-0003-1070-213X
Armin Scheben https://orcid.org/0000-0002-2230-2013
Douglas E. Soltis https://orcid.org/0000-0001-8638-4137
Pamela S. Soltis https://orcid.org/0000-0001-9310-8659
Soodeh Tirnaz https://orcid.org/0000-0002-6100-1790

Data availability

All sequencing data used in this study are available on SRA through accession numbers PRJNA212863, PRJEB4921 and PRJNA543572. The assembled genomes and other data are available at 10.26182/2ewa-pd24. Alignments of all individuals are available at 10.26182/rkcy-qa46. The genome can be visualised using JBrowse (Buels *et al.*, 2016) at appliedbioinformatics.com.au/amborella. The instance contains tracks displaying PAV displayed as pie charts in which red areas stand for lost genes and green areas stand for retained genes, SNPs and predicted genes. For the predicted genes, there are tracks showing the transcript evidence, the evidence from protein alignments and the evidence from *ab initio* gene predictors. The read depth for data from each individual is also displayed.

Haifei Hu[1], Armin Scheben[1,2], Brent Verpaalen[1], Soodeh Tirnaz[1], Philipp E. Bayer[1], Richard G. J. Hodel[3], Jacqueline Batley[1], Douglas E. Soltis[4,5,6,7], Pamela S. Soltis[5,6,7] and David Edwards[1]*

[1]School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA 6009, Australia;
[2]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA;
[3]Department of Botany, National Museum of Natural History, Smithsonian Institution, Washington, DC 20013-7012, USA;
[4]Department of Biology, University of Florida, Gainesville, FL 32611, USA;
[5]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA;
[6]The Genetics Institute, University of Florida, Gainesville, FL 32610, USA;
[7]The Biodiversity Institute, University of Florida, Gainesville, FL 32611, USA
(*Author for correspondence: email dave.edwards@uwa.edu.au)

References

Alexa A, Rahnenfuhrer J. 2010. *topGO: enrichment analysis for gene ontology.* R package v.2.22.0. doi: 10.18129/B9.bioc.topGO [accessed 19 July 2019].

Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* **342**: 1241089.

Andréfouët S, Torres-Pulliza D, Dosdane M, Kranenburg C, Murch B. 2004. *Atlas des récifs coralliens de Nouvelle-Calédonie.* Nouméa, New Caledonia: IRD.

Bayer PE, Edwards D, Batley J. 2018. Bias in resistance gene prediction due to repeat masking. *Nature Plants* **4**: 762–765.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nature Plants* **6**: 914–920.

Bayer PE, Golicz AA, Tirnaz S, Chan CK, Edwards D, Batley J. 2019. Variation in abundance of predicted resistance genes in the *Brassica oleracea* pangenome. *Plant Biotechnology Journal* **17**: 789–800.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L *et al.* 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* **17**: 66.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Cingolani P, Platts A, le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. 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* **6**: 80–92.

Conesa A, Gotz S. 2008. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics* **2008**: 1–12.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940.

Danecek P, Auton A, Abecasis G, Albers Ca, Banks E, DePristo Ma, Handsaker Re, Lunter G, Marth Gt, Sherry St *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Dolatabadian A, Bayer PE, Tirnaz S, Hurgobin B, Edwards D, Batley J. 2020. Characterization of disease resistance genes in the *Brassica napus* pangenome reveals significant structural variation. *Plant Biotechnology Journal* **18**: 969–982.

Elzhov TV, Mullen KM, Spiess A-N, Bolker B. 2016 *minpack. lm: R interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK.* R package v.1.2-1. [WWW document] URL https://cran.r-project.org/web/packages/minpack.lm/index.html [accessed 15 June 2019].

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL *et al.* 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature Genetics* **51**: 1044–1051.

Golicz AA, Batley J, Edwards D. 2016a. Towards plant pangenomics. *Plant Biotechnology Journal* **14**: 1099–1105.

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim HyeRan, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP *et al.* 2016b. The pangenome

of an agronomically important crop plant *Brassica oleracea*. *Nature Communications* 7: 13390.

Golicz AA, Martinez PA, Zander M, Patel DA, Van De Wouw AP, Visendi P, Fitzgerald TL, Edwards D, Batley J. 2015. Gene loss in the fungal canola pathogen *Leptosphaeria maculans*. *Functional & Integrative Genomics* **15**: 189–196.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L *et al.* 2017. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications* **8**: 2184.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**: R7.

Hampe A, Petit RJ. 2005. Conserving biodiversity under climate change: the rear edge matters. *Ecology Letter* **8**: 461–467.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K *et al.* 2014. Insights into the Maize Pan-genome and pan-transcriptome. *Plant Cell* **26**: 121–135.

Hodel RG, Chandler LM, Fahrenkrog AM, Kirst M, Gitzendanner MA, Soltis DE, Soltis PS. 2018. Linking genome signatures of selection and adaptation in non-model plants: exploring potential and limitations in the angiosperm *Amborella*. *Current Opinion in Plant Biology* **42**: 81–89.

Hurgobin B, Golicz AA, Bayer PE, Chan CKK, Tirnaz S, Dolatabadian A, Schiessl SV, Samans B, Montenegro JD, Parkin IA *et al.* 2018. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal* **16**: 1265–1274.

Jaffre T, Pillon Y, Thomine S, Merlot S. 2013. The metal hyperaccumulators from New Caledonia can broaden our understanding of nickel accumulation in plants. *Frontiers in Plant Science* **4**: 279.

Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, Xie W, Wang G, Yan J. 2016. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Scientific Reports* **6**: 18936.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li P, Quan X, Jia G, Xiao J, Cloutier S, You FM. 2016. RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**: 852.

Li Y-H, Zhou G, Ma J, Jiang W, Jin L-G, Zhang Z, Guo Y, Zhang J, Sui Yi, Zheng L *et al.* 2014. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature Biotechnology* **32**: 1045–1052.

Lillie A, Brothers R. 1970. The geology of New Caledonia. *New Zealand Journal of Geology and Geophysics* **13**: 145–183.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou G-A, Zhang H, Liu Z, Shi M *et al.* 2020. Pan-genome of wild and cultivated soybeans. *Cell* **182**: 162–176.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee HueyTyng, Chan C-K, Visendi P, Lai K, Doležel J, Batley J *et al.* 2017. The pangenome of hexaploid bread wheat. *The Plant Journal* **90**: 1007–1013.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**: 268–274.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867–868.

Poncet V, Munoz F, Munzinger J, Pillon Y, Gomez C, Couderc M, Tranchant-Dubreuil C, Hamon S, de Kochko A. 2013. Phylogeography and niche modelling of the relict plant *Amborella trichopoda* (Amborellaceae) reveal multiple Pleistocene refugia in New Caledonia. *Molecular Ecology* **22**: 6163–6178.

Rijzaani H, Bayer PE, Rouard M, Doležel J, Batley J, Edwards D. 2021. The pangenome of banana highlights differences between genera and genomes. *Plant Genome*. doi: 10.1002/tpg2.20100.

Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325–2329.

Seo E, Kim S, Yeom SI, Choi D. 2016. genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among solanaceae plants. *Frontiers in Plant Science* **7**: 1205.

Shao ZQ, Xue JY, Wu P, Zhang YM, Wu Y, Hang YY, Wang B, Chen JQ. 2016. Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat genes reveal three anciently diverged classes with distinct evolutionary patterns. *Plant Physiology* **170**: 2095–2109.

Shiu SH, Karlowski WM, Pan RS, Tzeng YH, Mayer KFX, Li WH. 2004. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**: 1220–1234.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**: 62.

Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M *et al.* 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**: aab3761.

Tan SJ, Zhong Y, Hou H, Yang SH, Tian DC. 2012. Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biology* **12**: 86.

Tang S, Lomsadze A, Borodovsky M. 2015. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research* **43**: e78.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**: 11–14.

Teurlai M, Menkès CE, Cavarero V, Degallier N, Descloux E, Grangeon J-P, Guillaumot L, Libourel T, Lucio PS, Mathieu-Daudé F *et al.* 2015. Socio-economic and climate factors associated with dengue fever spatial heterogeneity: a worked example in New Caledonia. *PLoS Neglected Tropical Diseases* **9**: e0004211.

Thien LB, Sage TL, Jaffre T, Bernhardt P, Pontieri V, Weston PH, Malloch D, Azuma H, Graham SW, McPherson MA *et al.* 2003. The population structure and floral biology of Amborella trichopoda (Amborellaceae). *Annals of the Missouri Botanical Garden* **90**: 466–490.

Torkamaneh D, Lemay MA, Belzile F. 2021. The Pan-genome of the Cultivated Soybean (PanSoy) reveals an extraordinarily conserved gene content. *Plant Biotechnology Journal*. doi: 10.1111/pbi.13600.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F *et al.* 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**: 43–49.

Zhang Y, Xia R, Kuang H, Meyers BC. 2016. The diversification of plant NBS-LRR defense genes directs the evolution of MicroRNAs that target them. *Molecular Biology and Evolution* **33**: 2692–2705.

Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Pan-genome modelling of *Amborella*.

Fig. S2 Abiotic stress-associated genes in *Amborella*.

Fig. S3 Plots show the number of presenting gene copies across populations in the gene family with Myb-like DNA-binding domain (PF00249).

Fig. S4 PAVs distribution and the clustering of the dispensable genes annotated as cadmium ion response-associated genes.

Fig. S5 Resistance gene families in *Amborella*.

Fig. S6 Venn diagram of the RGAs reported by our study and Shao *et al*. (2016) and Zhang *et al*. (2016).

Fig. S7 Upset plot of gene PAVs distribution among 11 *Amborella* samples.

Fig. S8 Protein domain analysis for the resistance gene analogue evm_27.model.AmTr_v1.0_scaffold00010.352.

Fig. S9 Protein domain analysis for the resistance gene analogue evm_27.model.AmTr_v1.0_scaffold00062.76.

Fig. S10 Manually checking the protein domains for the resistance gene analogues.

Table S1 Assembly statistics of newly assembled contigs in *Amborella* pan-genome and other pan-genome studies.

Table S2 Total number of core genes and dispensable genes in *Amborella*.

Table S3 GO enrichment of dispensable genes.

Table S4 List of *Amborella* dispensable genes.

Table S5 Abiotic stress-associated genes in *Amborella*.

Table S6 Number and percentage of dispensable genes in *Amborella trichopoda*, *Brassica oleracea*, *Glycine max*, *Brachypodium distachyon*, *Oryza sativa* L. and *S. lycopersicum* L.

Table S7 The gene family classification of cadmium responses associated genes.

Table S8 Resistance gene families in *Amborella*.

Table S9 Physical clustering of RGAs in *Amborella*.

Table S10 Comparison of the resistance gene analogue (RGA) classification in Shao *et al*. (2016) and classification used in our study based on RGAugury.

Table S11 Comparison of the resistance gene analogue (RGA) classification in Zhang *et al*. (2016) and classification used in our study based on RGAugury.

Table S12 Sequence similarity comparison between the TNL protein coding sequences (*A. trichopoda* v.1.0) used in Zhang *et al*. (2016) and sequences in our study.

Table S13 Gene PAVs statistics for different *Amborella* individuals.

Table S14 Geographical details of the 11 *Amborella* samples used in this study and the summary statistics of Illumina re-sequencing data per sample.

Table S15 Details of the RNA-seq libraries used in this study.

Table S16 Genes and sequencing reads in Mb for each line.

Table S17 Comparison of the conflict resistance gene analogue (RGA) classification in Zhang *et al*. (2016), Shao *et al*. (2016) and classification used in our study based on RGAugury.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Key words: abiotic stress, adaptation, *Amborella*, diversity, pan-genome.

Received, 18 June 2021; accepted, 26 July 2021.