Mapping the root systems of individual trees in a natural community using genotyping-by-sequencing

Osborne, Owen; Dobreva, Mariya P; Papadopulos, Alexander S. T.; de Moura, Magna S.B. ; Brunello, Alexandre T. ; de Queiroz, Luciano P. ; Pennington, R. Toby; Lloyd, Jon; Savolainen, Vincent

New Phytologist

DOI:
10.1111/nph.18645

Published: 01/05/2023

Peer reviewed version

Title:

Mapping the root systems of individual trees in a natural community using genotyping-by-sequencing

Article type

Methods Paper

Authors

Owen G. Osborne[1,2,*], Mariya P. Dobreva[1], Alexander S.T. Papadopulos[2], Magna S.B. de Moura[3], Alexandre T. Brunello[4], Luciano P. de Queiroz[5], R. Toby Pennington[6,7], Jon Lloyd[1], Vincent Savolainen[1,8,*]

Author affiliations

[1] Georgina Mace Centre for the Living Planet, Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot, SL5 7PY, UK

[2] Molecular Ecology and Evolution Bangor, School of Natural Sciences, Bangor University, Environment Centre Wales, Deiniol Road, Bangor, LL57 2UW, UK

[3] Empresa Brasileira de Pesquisa Agropecuária, Petrolina, PE, Brazil

[4] Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Departamento de Biologia, Universidade de São Paulo, Av. Bandeirantes, 3900, Monte Alegre, 14040-901, Ribeirão Preto, SP, Brazil

[5] Universidade Estadual de Feira de Santana, Dept. Ciências Biológicas. Av. Transnordestina s.n., Novo Horizonte. 44036-900, Feira de Santana, BA, Brazil

[6] Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX4 4RJ UK

[7]Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, EH3 5LR, UK

[8]Royal Botanic Gardens, Kew, Richmond, TW9 3AB, UK

*authors for correspondence: v.savolainen@imperial.ac.uk and owengosborne@gmail.com.

Summary

- The architecture of root systems is an important driver of plant fitness, competition and ecosystem processes. However, the methodological difficulty of mapping roots hampers the study of these processes. Existing approaches to match individual plants to belowground samples are low-throughput and species-specific. Here, we developed a scalable sequencing-based method to map the root systems of individual trees across multiple species. We successfully applied it to a tropical dry forest community in the Brazilian Caatinga containing 14 species.
- We sequenced all 42 individual shrubs and trees in a 14 by 14 m plot using double-digest restriction-site associated sequencing (ddRADseq). We identified species-specific markers and individual-specific haplotypes from the data. We matched these markers to ddRADseq data from 100 mixed root samples from across the centre (10 by 10 m) of the plot at four different depths, using a newly developed R package.
- We identified individual root samples for all species and all but one individual. There was a strong significant correlation between below and aboveground size measurements, and we also detected significant species-level root-depth preference for two species.
- The method is more scalable and less labour-intensive than current techniques, and is broadly applicable to ecology, forestry and agricultural biology.

Keywords:

Individual root density distribution; Belowground plant ecology; Caatinga; ddRADseq; Tropical community

Introduction

Most plant ecology studies have focussed on aboveground traits, despite a large proportion of plant biomass being located belowground (Mokany *et al.*, 2006; Poorter *et al.*, 2012). This has led to limited research into crucial processes occurring in the soil, such as plant–soil, plant–microbial and plant–plant interactions and their implications for ecosystem processes (Bardgett *et al.*, 2014). Expanding our knowledge in this area has implications for biodiversity conservation, plant productivity, and predicting ecosystem responses to global environmental change (Ostle *et al.*, 2009).

Assessing root distribution at the individual level permits a reconstruction of the fine rooting patterns of single plants (e.g., individual trees in an area of a forest) in three dimensions. This then allows inferences of how plant roots compete with each other for nutrients and water, and the relationship between aboveground and belowground biomass — contributing to the understanding of the structure and dynamics of community-level and evolutionary processes such as niche differentiation, symbiosis and environment-phenotype interactions. To achieve this, better methodologies are needed for detecting the distribution of individual root systems.

Belowground studies in natural systems are limited by the difficulty of observing roots in natural settings, which is especially true for trees where excavation of entire root systems is destructive and sometimes unfeasible (Cabal *et al.*, 2021). Therefore, alternative techniques are needed to elucidate the belowground structure and interactions of particular plant species, or ideally, specific individuals (Jones *et al.*, 2011; Cabal *et al.*, 2021). Methods based on DNA sequencing and related computational techniques have allowed an increasing number of assessments of belowground plant distribution at the species level (Jackson et al., 1999; Bardgett et al., 2014). To differentiate roots of different species, amplicon sequences are usually sequenced in mixed root DNA from soil cores and are then allocated to species by comparison to databases (Mommer *et al.*, 2010; Bardgett *et al.*, 2014; Barberán *et al.*, 2015). DNA metabarcoding has been successfully used to identify the species composition (Jones *et al.*, 2011; Kesanakurti *et al.*, 2011; Hiiesalu *et al.*, 2012) and relative abundance (Matesanz *et al.*, 2019) of plant communities from mixed root samples. However, this

approach is successful for species-level identification only, and is dependent on the existence of complete reference libraries (Jones *et al.*, 2011).

Microsatellite markers have been used to assign single root fragments to individual trees (Saari *et al.*, 2005). This approach, however, is not applicable to experiments with large sample sizes since it requires each root fragment to be processed individually, which is laborious. Furthermore, species-specific PCR primers for each marker must first be developed in order to use microsatellite approaches (Zane *et al.*, 2002), limiting their scalability to mixed plant communities. To the best of our knowledge, no high-throughput method has been successful in linking root DNA from mixed-species soil specimens to individual plants.

The restriction-site associated DNA sequencing family of methods (RADseq; also known as Genotyping-by-Sequencing, Davey and Blaxter 2010) have been employed to address a wide variety of ecological, phylogenetic and evolutionary questions (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). These include resolving relationships among closely related species (Grewe *et al.*, 2017), tracing the movement of insects among host plants (Fu *et al.*, 2017), population genetic inference of selection (Magalhaes *et al.*, 2020) and building genetic maps (Papadopulos *et al.*, 2019). The double digest variation of the RADseq method (ddRADseq) can be used for Single Nucleotide Polymorphism (SNP) discovery and genotyping of any organism, without the need of a reference genome (Heyland & Hodin, 2004; Peterson *et al.*, 2012; Andrews *et al.*, 2016). This makes ddRADseq a relatively inexpensive and potentially suitable approach for tracing individual plant roots from mixed soil samples.

Here, we describe a method to allow direct inferences on the fine rooting patterns of individual trees. We employed ddRADseq data from all individual trees, a single specimen of each shrub species present, and 100 mixed root samples, across an experimental plot in the understudied but ecologically important seasonally dry tropical forest of the Brazilian Caatinga. We developed a bioinformatic pipeline to link the individual trees to root samples using this data, constructed 3D maps of fine root distribution of each tree, and used the results to identify species-specific root-depth niches and aboveground-belowground size correlations.

Materials and Methods

Study design and sequencing

Our study site consisted of a 14 × 14 m plot situated on the Semiarid unit of the Brazilian Agriculture Research Corporation (EMBRAPA; Pernambuco State, Brazil; central coordinates: 9.04002°S, 40.31957°W; Fig. 1a). The studied vegetation can be broadly described as being part of the Caatinga domain (de Lima Araújo *et al.*, 2007) with soil physical and chemical properties sampled and analysed as in Quesada *et al.* (2011) yielding a World Reference Base (IUCC, 2006) soil classification of “Haplic Lixisol (Loamic, Hypereutric, Ochric, Magnesic)”. The Brazilian Caatinga is recognised as the largest and most species rich forests of the Seasonally Dry Tropical Forest (SDTF) biome in the New World (Pennington *et al.*, 2000, Fernandes *et al.,* 2022).

To quantify the vegetation structure, measurements of stem diameters and projected canopy areas were made according to protocols as detailed in Tortello-Raventos et al. (2013) and Moonlight et al. (2021). Tree height measurements were taken by holding a graduated pole close to the trunk. Tree height and crown base height correspond to the distance from ground level to the highest and lowest fully expanded leaf, respectively. The main stem diameter at breast height (1.3 m; DBH) and the visible crown extension in two cardinal directions were measured. The canopy volume was calculated assuming an ellipsoid shape and canopy area was calculated assuming an elliptical shape (Sampaio and Silva, 2016). The few subshrub and succulent herbaceous species were not measured.

This yielded estimates (all woody plants with a stem DBH > 25 mm) of a stem density of *ca.* 2420 ha^{-1}, a woody plant canopy area index of 1.39 m^2 m^{-2}, and with mean and 0.95 quantile canopy heights of 3.9 m and 7.5 m respectively. Although there was also a subordinate herbaceous and shrub understorey present, this visually estimated to be with a total fractional cover of much less than 0.3. This, along with the clear dry-deciduous nature of the majority of the species present allowed the studied vegetation to be classified as a ‘closed deciduous shrubland’ (Torello-Raventos et al., 2013).

The study stand consisted of both trees and shrubs (as defined by Tortello-Raventos *et al.,* 2013) with all 42 woody individuals of DBH > 25 mm present sampled for DNA extraction. This woody component consisted of *Cenostigma microphyllum* (Mart. ex G. Don) Gagnon & G.P. Lewis, *Cereus albicaulis* (Britton & Rose) Luetzelb., *Chloroleucon foliolosum (Benth.)* G.P. Lewis, *Cnidoscolus quercifolius* Pohl, *Commiphora leptophloeos (Mart.) J.B. Gillett, Croton echioides* Baill.*, Handroanthus spongiosus* (Rizzini) S.O. Grose, *Jatropha mollissima* (Pohl) Baill., *Manihot carthagenensis (Jacq.) Müll. Arg.*, *Mimosa arenosa* (Willd.) Poir., *Pseudobombax simplicifolium* A. Robyns, *Sapium glandulosum* (L.) Morong, *Schinopsis brasiliensis* Engl., and *Senna macranthera (DC. ex Collad.) H.S. Irwin & Barneby*. We also sampled one individual of each of the five subshrub and succulent herbaceous species present, these being *Calliandra depauperata* Benth., *Ditaxis desertorum* (Müll. Arg.) Pax & K. Hoffm., *Neoglaziovia variegata* (Arruda) Mez, *Tacinga inamoena* (K. Schum.) N.P. Taylor & Stuppy, and *Varronia leucocephala (Moric.) J.S. Mill*. This resulted in 47 aboveground samples in total (Table S1). For each specimen, we collected fresh leaf samples with an approximate size of 4 cm^2, which were cut into 3 mm strips and stored in RNA*later* (Sigma) until further processing.

For root collection, the centre (10 × 10 m) of the plot was subdivided into a grid of 2 × 2 m subplots (Fig. 1b). Soil cores were sampled from the centre of each subplot with an auger with a core of 6.25 cm diameter. Four samples, representing four different depth ranges (0–5 cm, 5–10 cm, 15–20 cm, and 45–50 cm; Fig. 1c), were then taken from each core for root sampling, resulting in 100 root samples in total. All roots within each core-sample were separated from the soil in the field with a metal sieve, washed with water, and preserved in RNA*later* (Sigma) until further processing.

The 100 mixed root samples and 47 leaf samples were sent to LGC (Berlin, Germany) for DNA extraction, library construction, and sequencing. Mixed root samples were homogenised prior to DNA extraction such that aliquots used for extraction were likely to contain a mixture of all roots in the entire sample. Approximately 100 mg of homogenised root or leaf material was used to extract DNA from each sample using a

CTAB-chloroform method (Xin and Chen, 2012). Illumina paired-end (2 × 150 bp) double-digest restriction-associated DNA libraries were prepared using *pst*I and *ape*KI restriction enzymes (Hamblin & Rabbi, 2014) and were sequenced on an Illumina NextSeq 550 machine.

Bioinformatic pipeline

We developed a pipeline to use species-specific ddRAD markers and individual-specific haplotypes of these markers, to link species and individual trees present in a plot to unknown root samples collected from the soil below it. Our method uses the *STACKS* pipeline (version 2.52; Rochette et al. 2019) as well as a new *R* package – *RootID* (version 1.0). The pipeline follows three steps: i) Generate a catalogue of all markers and haplotypes across all leaf samples and match all data from both leaf and root samples to it, (*STACKS)*; ii) Identify diagnostic markers and haplotypes from the leaf catalogue: those which are unique to a species or an individual, (*RootID)*; iii) Match the root data to these diagnostic markers or haplotypes, to determine which are found in which root samples, and thus which tree's roots are likely to be present in them (*RootID)*.

Sequence data was first demultiplexed, adaptor sequences and Illumina barcodes were clipped and reads were filtered to ensure they contained the correct restriction enzyme cut sites by LGC using their in-house pipeline. We input these reads into *process_radtags* module from *STACKS* to further filter them and prepare them for the main *STACKS* pipeline. We used the *c* option to remove any reads with uncalled bases, trimmed reads to exactly 142 bases in length (the expected read-length with adaptor sequences removed; *t* = 142, *len-limit* = 142), and removed reads where the PHRED-scaled quality score fell below 25 in a sliding window of 15% of read length (*q*, *s* = 25, *w* = 0.15). Processed read pairs were then concatenated, without merging of overlapping sequence. This is justified, since *STACKS* requires sequences of the same length, and our downstream analyses identify sequences based on exact sequence identity (as opposed to, for example, genetic distance between sequences) so a portion of the sequence being repeated has no impact of assignment.

We ran the *ustacks* module on all (root and leaf) samples to build sample specific sets of loci. We used the deleveraging algorithm (*d*), disabled haplotype calling from secondary reads (*H*) and disabled gapped alignment between stacks *(disable-gapped*). We used a minimum depth (*m*) of 1 for root samples and 5 for leaf samples (n.b. more stringent depth filters were applied in the post-processing of *STACKS* output using our *RootID* package). We used multiple values for the *ustacks M* parameter, and selected the best using our optimisation procedure (see *Pipeline optimisation*, below). We then ran *cstacks* on the leaf samples to build catalogues, again disabling gapped alignment between stacks (*disable-gapped*). As with the *ustacks M* parameter, we used multiple values for the *cstacks n* parameter and selected the best (see below). We then matched all sets of loci built with *ustacks* to the catalogue using *sstacks* with gapped assembly turned off (*disable-gapped*), which produced the input files required for *RootID*.

RootID takes the *matches.tsv.gz* files produced by *sstacks* as input. One of these files is produced by *sstacks* for each sample, which contains the read depth for all haplotypes that were matched to the catalogue. The main workflow of the package is implemented in three functions: i) *read.stacks*, which reads *sstacks* output files for all known aboveground (leaf) samples and converts them to an *R* object; ii) *find.diag*, which identifies species-diagnostic loci and individual-diagnostic haplotypes from the output of *read.stacks*; and iii) *match.diag*, which matches these diagnostic markers and haplotypes to those in root samples.

The *find.diag* function first identifies species-diagnostic markers (i.e. putative genomic loci), those which are unique to a single species in the dataset. These must be absent in all heterospecific individuals (at any read depth) and must occur in a user-defined proportion of individuals of the focal species (optional thresholds of minimum read depth per individual and maximum number of haplotypes per marker can also be applied). The function then identifies individual-diagnostic haplotype variants within the species-specific markers that are unique to a single individual.

Matching to the root samples is achieved with the third function in the pipeline, *match.diag*, which reads all *sstacks* output files for root samples, matches them to the diagnostic markers and haplotypes identified by *find.diag*, and reports the number of

reads in each root sample that match diagnostic markers and haplotypes for each species and individual tree. For the analyses presented in the main text, we considered there to be a match if any diagnostic markers or haplotypes were detected in the root samples because false positives are likely to be far less frequent than false negatives (see discussion). However, *match.diag* can optionally filter matches by a minimum number of reads, and we also present results using a minimum read number of 3 in the supplementary information. Doing so will likely increase specificity at the cost of reducing sensitivity.

Pipeline optimisation

We ran the data through our pipeline in multiple runs where we varied several important parameters to determine their effect on the results. The maximum number of mismatches allowed between alleles to merge them into a putative locus (*M*) is one of the main parameters that affect the level of polymorphism in *STACKS* (Paris *et al.*, 2017), so we ran *ustacks* separately with a range of values of *M*: 2, 4, 6 and 8. Previous work showed that the optimal number of mismatches allowed between putative loci when building the catalogue (*n*) was between $M - 1$ and $M + 1$ (Paris *et al.*, 2017).

Therefore, for each value of *M* used in *ustacks*, we built a catalogue with $n = M - 1$, $n = M$, and $n = M + 1$, resulting in 12 catalogues overall. When we ran *sstacks*, we matched each set of *ustacks* outputs (i.e. *ustacks M* = 2, 4, 6 or 8), to the three catalogues that were produced with the same value of *M* (i.e. *cstacks n* = *M* -1, *M*, or *M* + 1), resulting in 12 sets of *sstacks* results.

We ran the *RootID* pipeline on each of the 12 sets of matches produced by *STACKS*, separately. There are several parameters in the *find.diag* function that have the potential to affect the results, so we used a range of values for each of these. For the *min.dep* parameter, which sets the minimum read depth required for a marker to be considered present in an individual, we used values of 5, 10 and 20. For *max.md.marker*, which controls the maximum proportion of missing data among individuals of the focal species to call a species-diagnostic marker, we used values of 0, 0.2 and 0.4. For *max.md.hap* which controls the maximum proportion of missing data to

call an individual-diagnostic haplotype, we used values of 0, 0.1 and 0.2 (with 0 only used when *max.md.marker* was also set to 0). For *max.haps*, which controls the maximum number of haplotypes allowed per-marker, we used values of 2, 3 and *NA*, where *NA* specifies no limit. Using every combination of *STACKS* and *RootID* parameters resulted in 756 sets of results. To choose the optimal set of parameters, we ranked the results by number of individual-diagnostic haplotypes for each individual, and chose the results with the best mean rank for downstream analyses.

To assess the robustness of our results to parameter choice, we compared the results when each set of parameters (i.e., *STACKS* settings, minimum sequence depth, maximum missing data and maximum haplotypes) was varied while all other parameters were fixed at the optimal values identified above.

Pipeline validation

To confirm the effectiveness of using marker presence or absence to distinguish the species present in the plot, we used a hierarchical clustering approach. We first constructed a matrix of the proportion of shared markers across all individuals (i.e., the proportion of the markers in the individual with fewer markers that are shared with the individual with more markers). This was then used to calculate an unweighted pair group method with arithmetic mean (UPGMA) dendrogram using the function *upgma* in the *R* package *phangorn* (v. 2.5.5; Schliep 2011) in *R*. If marker presence is an effective method to distinguish species, conspecific individuals should cluster monophyletically in the resulting dendrogram. We include a function to conduct this analysis, *shared.marker.tree*, in the *RootID* package.

We then assessed how thoroughly the diversity has been sampled in each root sample at the level of haplotype, marker, individual and species, using rarefaction analysis. By randomly subsampling the data across a range of subsample sizes, it is possible to estimate whether the sampling effort is sufficient to identify all diversity present in the total sample. If all diversity present (e.g. all species) has been detected using 50% of the data, for example, then the addition of the remaining 50% of data will not lead to an increase in detected diversity. Therefore, in the above example, if subsample size is plotted against detected diversity, the horizontal asymptote will be reached at around

50%. For each root sample, we randomly subsampled between 2% and 98% of the reads that matched our catalogue without replacement at 2% intervals. This was repeated 100 times for each rarefaction level and the mean and 95% quantile of number of unique species-diagnostic markers, individual-diagnostic haplotypes, individuals and species was calculated. We include a function, *sample.rarefaction*, in the *RootID* package to conduct this analysis. The results were used to plot rarefaction curves for each root sample. We calculated the slope of the final 10% of the curve as:

$$m = \frac{1 - P_{90}}{0.1}$$

Where P_{90} is the mean proportion of total diversity detected at 90% rarefaction (i.e., the mean proportion of species-diagnostic markers, individual-diagnostic haplotypes, species or individuals detected using the whole dataset that were detected when 90% of the data was randomly subsampled). Values closer to zero indicate higher sufficiency of sequencing effort. We tested whether variation in *m* was correlated with the number of sequenced reads per root sample using Spearman's correlation tests.

We expected that the roots of each tree would be more likely to be found in samples located closer to the tree and, if the method worked well, this would be reflected in the results. To test this expectation, we first calculated the Euclidian distance (ignoring root sample depth) between tree and root sample locations for all tree and root sample pairs for which the individual tree was detected in the root sample (matches). We compared this to the distance between all tree and root sample pairs for which the tree was not detected in the root sample (non-matches) using Mann-Whitney U tests. We considered significantly lower distance in matches than non-matches as evidence that tree roots are more likely to be detected in samples closer to the tree. We took a similar approach to the same question using our species-diagnostic marker results, but took the distance from the root sample to the nearest tree of the focal species for species with multiple individuals.

Finally, we used simulated data to assess the effect of genome size and sequencing depth on the number of diagnostic markers and haplotypes recovered. We downloaded six genome assemblies from Phytozome (Goodstein et al., 2012): *Arabidopsis thaliana*

(version: Araport11; total scaffold length: 120 Mbp), *Populus trichocarpa* (version: 4.1; 392 Mbp), *Eucalyptus grandis* (version: 2.0; 691 Mbp), *Asparagus officinalis* (version: 1.1; 1,188 Mbp), *Lactuca sativa* (version: 8; 2,400 Mbp) and *Helianthus annuus* (version: r1.2; 3,028 Mbp). These were used to generate simulated ddRAD reads using *RADinitio* (Rivera-Colón et al., 2021). We first simulated 10 individuals of each species using the *make-population* command in *RADinitio* using a simulated population size of 1,000. The simulated individuals were then used to simulate ~1,000,000 read pairs per species (using the appropriate *-coverage* setting for the genome size of each species) using the *make-library-seq* command in *RADinitio* with 10 simulated PCR cycles, a read length of 150, and the enzymes *PstI* and *MspI* (*ApeKI* is not available in *RADinitio*). The simulated reads were then randomly subsampled between 100,000 and 1,000,000 reads (with a step-size of 100,000) using the *sample* command in *seqtk* (https://github.com/lh3/seqtk; version 1.3-r117-dirty) with random seeds recorded to ensure reproducibility (Table S7). Subsampled reads were then processed with *ustacks*, *cstacks*, *sstacks*, *read.stacks* and *find.diag* using the optimal setting identified above. The results of *find.diag* were used to plot the relationship between number of reads and numbers of diagnostic markers and haplotypes for each species and individual, respectively. Correlations between genome size and number of diagnostic markers and haplotypes were tested using Spearman's correlation tests.

Visualisation

We visualised the results in the form of three-dimensional root "maps" for each species and individual using a function, *plot_roots_3d*, in the *RootID* package. This uses the *rgl* package in *R* (Murdoch & Adler, 2021) to show the root sampling layout as a three-dimensional grid. Each grid square represents one root sample, and visually displays the abundance of the focal tree or species (either in the form of colour intensity or density of randomly distributed particles within each root sample). Optional three-dimensional models of the trees show their position, height, crown base height, and crown diameter. We used the *plot_roots_3d* function to produce root maps for all species and all individuals.

Analysis of the root distribution patterns

We used the results to detect broad belowground distribution patterns among the species in the plot.

Firstly, we asked whether the belowground distribution of each species was significantly associated with root-sample depth using linear-by-linear association tests in the *coin* package in *R* (Agresti, 2002; Hothorn et al., 2008) in each species separately. *P*-values were corrected for multiple-testing using the false discovery rate method (Benjamini & Hochberg, 1995).

Secondly, we asked whether the dimensions of the aboveground and belowground portions of the trees were correlated. We first calculated two belowground size metrics: i) the root radius, which we defined as the horizontal distance from each individual tree's trunk to the furthest root sample in which it was detected, and ii) the number of root samples each individual was detected in. We then tested whether these measures were significantly correlated with five aboveground size metrics: tree height, crown base height, canopy radius, canopy area and canopy volume. Because the number of individuals per species can reduce the number of potential individual-specific haplotypes, which in turn may reduce the chance that an individual is detected in any given root sample (see results), we used a partial Spearman's correlation test using the *pcor.test* function in the *R* package *ppcor* (Kim, 2015). This tested for correlation between root size and aboveground measurements while controlling for number of conspecific individuals.

Results

Matching roots to aboveground trees

The sequencing produced between 223,378 and 1,045,252 read-pairs for leaf samples and between 133,584 and 1,523,847 read-pairs for root samples following filtering (Table S2). Of the 756 parameter combinations tested, the optimal parameters for each pipeline component were as follows: for *ustacks*, *M* = 6; for *cstacks*, *n* = 7; for *find.diag*, *max.md.marker* = 0.4, *max.md.hap* = 0.2, *min.dep* = 5, and *max.haps* = unlimited (Table S3). The results produced using the optimal parameter combination were used for all subsequent analyses.

The leaf data were assembled into 316,537 catalogue loci across all individuals. Using the *read.stacks* and *find.diag* functions in *RootID*, between 6,842 and 16,814 species-specific markers were identified per species (Table S3). Diagnostic haplotypes were identified for all individuals, but these varied in number from 10 to 7,420 per individual (Table S3).

Using the *match.diag* function, between 67 and 91,223 root reads per sample were mapped to catalogue markers. Of these, between 14.49% and 99.94% were matched to species-diagnostic markers, and between 0% and 25.78% were matched to individual-specific haplotypes (with 91/100 root samples having at least one match to an individual-specific haplotype; Table S4).

All 14 tree/shrub species were detected in between 5 and 90 of the 100 root samples and all 5 subshrub/herb species were detected in between 26 and 94 root samples (Fig. 2a and c; Figs. S1 – S17). Of the 37 individuals (i.e., those from species with multiple individuals for which the individual-specific haplotype analysis was conducted), 36 were detected in at least one root sample (median = 10 root samples; Fig. 2b and d; Figs. S18–S24). The undetected individual (L_22) was from the species with the fewest individual-diagnostic haplotypes, *Jatropha mollissima*.

Patterns of root distribution

We found that two species, *Cenostigma microphyllum* and *Ditaxis desertorum,* had depth distributions that significantly departed from null expectations following multiple test correction (Fig. 3; when a minimum read depth filter of 3 was used in *match.diag* (see methods), *D. desertorum* no longer had a significant association with depth but an additional species: *Varronia leucocephala* did; Table S5). Both species were more commonly detected in the two deeper root depth levels (15–20 cm and 45–50 cm) than at shallower levels. Lateral aboveground size metrics (canopy radius, canopy area and canopy volume) were significantly positively correlated with the number of root samples each individual was detected in, while controlling for number of individuals per species (Spearman's partial correlations: canopy radius: $\rho = 0.58$, $P = 0.0001$; canopy area: $\rho = 0.59$, $P < 0.0001$; canopy volume: $\rho = 0.44$, $P = 0.006$; Fig. 4). The correlation between number of root samples and tree height was marginally non-significant ($\rho = 0.31$, $P =$

0.053). In contrast, there was no significant correlation between root radius and any aboveground metrics (Table S6). When a minimum depth filter of 3 was used for *match.diag* (see methods), results were similar in terms of significance/non-significance except for the correlation between number of root samples and tree height, which was significant with this filter (Table S6).

Pipeline optimisation and validation

The parameter comparison showed that the analysis was fairly robust to the choice of parameter values. For the *STACKS* parameters, 94% of root-to-species and 74% of root-to-individual matches were found across all parameters values; for *max.md.*marker and *max.md.hap*, 99% of root-to-species and 97% of root-to-individual matches were found across all parameters values; for *min.dep*, 90% of root-to-species and 95% of root-to-individual matches were found across all parameters values; and for *max.haps*, 96% of root-to-species and 77% of root-to-individual matches were found across all parameters values (Figs. S25–S32). The pipeline was computationally efficient and did not require high-performance computing capabilities: the *RootID* analysis completed in between 96 and 114 seconds per run, on an *Apple Macbook Pro* laptop computer (16 GB memory) using a single processor.

Identification of both species-specific markers and individual-specific haplotypes was more efficient in species with fewer individuals. While this negative relationship was moderate for species-specific markers (Spearman's correlation test: $\rho = -0.47$, P = 0.04), it was strong and highly significant for individual-specific markers (Spearman's correlation test: $\rho = -0.62$, $P < 0.0001$). In our UPGMA clustering analysis based on the proportion of shared markers between individuals, all conspecific individuals clustered monophyletically, supporting the use of presence or absence of RAD markers for species identification (Fig. S33).

Individuals were more frequently detected in root samples that were physically closer to them (Mann-Whitney *U* test. W = 575509; $P < 0.0001$; Fig. S34) and species were more frequently detected in root samples that were closer to an individual of that species (Mann-Whitney U tests. W = 174844; $P < 0.0001$; Fig. S35). As with the number of

diagnostic haplotypes and markers (above), there was a significant negative correlation between the number of root samples an individual was detected in and the number of individuals per species, but there was no such correlation for species (Spearman's correlation tests. Species: ρ = 0.17, P = 0.47; individuals: ρ = -0.34, P = 0.04; Fig. S26).

The rarefaction analysis showed that final 10% slopes were high for species-diagnostic markers (median m = 0.58; Figs. S36–S39) and individual-diagnostic haplotypes (median m = 0.58; Figs. S40–S43). No root samples had m = 0 for either species-diagnostic markers or individual-diagnostic haplotypes. The number of reads per sample was significantly negatively correlated with final slope for both species-diagnostic markers and individual-diagnostic haplotypes (Spearman's correlation tests. Markers: $\rho = -0.69$, P < 0.0001; haplotypes: $\rho = -0.69$, P < 0.0001; Fig. S44). The final slopes for species (median m = 0.25; Figs. S45–S48) and individuals (median m = 0.22; Figs S49–S52) were much lower on average, and were zero for several samples (3 for species and 26 for individuals). In contrast to the results for markers and haplotypes, there was no significant correlation between the number of reads and the final slope of either species or individuals (Spearman's correlation tests: Species: $\rho = -0.18$, P = 0.08; individuals: ρ = 0.09, P = 0.39; Fig. S44).

The simulated data analysis showed that increased read depth increases the number of both species-diagnostic markers (Fig. S53a) and individual-diagnostic haplotypes (Fig. S53b). However, for most species, the majority of diagnostic markers and haplotypes are identified at relatively low sequencing depths. The number of individual-diagnostic haplotypes significantly increased with genome size (Fig. S53d; Spearman's correlation test: ρ = 0.94, P = 0.005). There was no association between genome size and number of species-diagnostic markers, however (Fig. S53c; ρ = -0.71, P = 0.11).

Discussion

Given the limitations of previous methods to genetically identify and map tree roots (i.e. DNA barcoding is appropriate only for species-level and microsatellites need species-specific developments), we designed here a new method, which has also been

validated by our dataset from the dry forest of Brazil. While the ideal control — a reliable spatial map of the fine roots in the plot by which to ground truth the results — is not feasible, the highly significant association between root position and tree position provides corroboration of the method (Fig. S34). The presence/absence of RAD loci is not usually treated as informative, but rather as missing data (Cerca *et al.*, 2021; Crotti *et al.*, 2019). This is largely because, while the presence or absence of a marker may result from mutational processes such as point mutations in the enzyme cut-site or indels which drastically alter fragment size, it can also result from technical issue in library preparation and sequencing (Cerca *et al.*, 2021). The rate of marker presence/absence variation from mutational processes is expected to increase with lineage divergence (Cerca *et al.*, 2021). Therefore, we expect that, in a dataset which includes multiple distantly related species such as ours, the majority of marker presence/absence variation is likely to be mutational rather than technical, and thus be useful for species differentiation. Indeed, our hierarchical clustering analysis (Fig. S33) indicates that marker presence/absence distinguishes species well in our dataset. However, since there were no congeneric species, it is possible that for closely related species this will be less effective. Therefore, we recommend that hierarchical clustering analysis should be performed in all cases, and species which cannot be reliably distinguished should be coded as a single species for the purpose of the analysis, such that individuals may still be distinguished using haplotype information. While we focussed on testing the method in a real dataset, future work could also evaluate the tolerance of the method for particularly closely related species using “pseudo-samples” – similar to the mock communities used as controls in metabarcoding analysis (Braukmann et al., 2019). This could be achieved by sequencing pairs of species with differing levels of relatedness to produce a catalogue, and making mixed pseudo-samples of known quantities of each of the species’ tissue, which could also be sequenced to test the limits and sensitivity of the method.

While analysis of RADseq data requires the selection of several parameters which can have large effects on downstream analysis, our results were highly robust to parameter choice. Furthermore, the computational efficiency of the pipeline allows many parameter combinations to be easily tested. False positive matches between individual trees and

root samples are likely to be relatively rare using our method, but may occur occasionally due to sequencing or PCR error. The chance of false positives is likely to be affected by multiple factors, including sequencing error rate and the number of SNPs distinguishing diagnostic haplotypes. However, it is worth noting that misidentification of individuals is very unlikely even with small numbers (~10) of unlinked and variable loci, a fact that forms the basis of forensic DNA fingerprinting (Norrgard, 2008). The false positive likelihood can be reduced by filtering the results of *match.diag* by a minimum number of markers or haplotypes (using the *min.reads.mar* and *min.reads.hap* options in *match.diag*, respectively), although this will likely increase the false negative rate. Here, we present both unfiltered matches (main text) and matches filtered by a minimum of 3 reads per match and find that while there were fewer matches in the filtered results, the overall findings of both the aboveground/belowground correlation and depth niche analysis were similar.

False negatives are likely to be much more common. The non-detection of an individual in a subplot could have one of several causes: i) they may be genuinely absent from the subplot; ii) they may be absent from the soil core taken to represent the sub-plot but present elsewhere in the subplot; iii) they may be present in the soil core, but the sequencing depth is insufficient to detect their diagnostic haplotypes. Since in our sampling regime each root sample is taken from a small fraction of the total volume of the subplot (153.4 cm^3 of a total 200,000 cm^3), it is likely that some trees present in some subplots were not captured by the soil-core sampling. This possibility is common to any soil core-based method and would be made less likely with denser sampling. Its likelihood may also be influenced by differences in root architecture between species, for example, it could be less common in species with a higher density of fine roots.

We estimated the sufficiency of our sequencing depth using a rarefaction-based approach similar to those employed in metabarcoding analyses (Estaki *et al.*, 2020). While none of the curves flattened at the marker and haplotype levels, several did at the individual and species levels. This indicates that while the sequencing effort was insufficient to sequence all diagnostic markers and haplotypes in the samples, this effect was substantially ameliorated at the level of species and individual detection

because there are multiple markers and haplotypes which can be used to detect each species or individual. Nevertheless, the success of the analysis was clearly linked to sequencing coverage and some samples performed poorly. The number of individuals per species was negatively correlated with both number of diagnostic markers and haplotypes detected in the roots, and the number of root samples an individual was detected in. This is expected: given a community of two individuals, all fixed genetic differences between them can be used as individual-diagnostic haplotypes to distinguish them. As more individuals are added to the community, there is a higher chance that another individual carries these haplotypes. This is likely to be exacerbated in populations with low genetic diversity, such as inbred populations, since they contain fewer intraspecific genetic variants overall. Sequencing effort also affects the number of diagnostic markers and haplotypes in the catalogue, as evidenced by our simulated data analysis. While none of the species in our Caatinga dataset have sequenced genomes, studies involving species with available genomic resources could make use of similar simulation studies to estimate the required sequencing depth prior to experimental design, significantly improving the efficiency and effectiveness of the approach. Thus, the number of identified diagnostic markers and haplotypes can be increased by higher sequencing depth in the aboveground tissues, and the number of these that are detected can be increased by higher sequencing depth in belowground samples. Both of these are likely to be more important if high numbers of conspecific individuals are present and in populations that are less genetically diverse. The impact of these caveats depends strongly on the research question. False negatives should be randomly distributed amongst samples. Therefore, even if detection capability differs among species, experiments addressing, for example, the vertical distribution of roots, are unlikely to be biased by this. Contrastingly, care should be taken if attempting to use these methods to compare absolute root biomass between species if they vary in number of individuals.

The analysis successfully identified species-diagnostic markers and individual-diagnostic haplotypes for all species and individuals and detected all species and all but one individual in root samples. Given that the total soil volume the roots were sampled from (0.015 m^3) was only 0.03% of the total volume of the plot (50 m^3), this implies that

the roots of most individuals are likely densely and widely distributed in the plot. Root distribution was variable between individuals and species, however. Number of root samples was significantly correlated with several measures of aboveground size. While not a direct measurement of root dimensions, number of root samples is likely to be influenced by both root system size and root density. There were no significant correlations between root radius and aboveground traits. Such a correlation has been shown in previous studies (Tumber-Dávila et al., 2022), and its absence here may be a result of many of the study plants extending their root systems beyond the bounds of the plot.

In this paper we have developed, to our knowledge, the first method capable of high-throughput individual-level root identification across multi-species plant communities. Given the fact that we were able to detect 97% of individuals across such a broad assemblage of plant species, the method is highly promising. It is also likely to be applicable to several distinct research questions. For species-level root identification, the current state of the art (metabarcoding) can suffer from lack of species differentiation at sequenced markers. This can be somewhat ameliorated by using multiple markers (Zhang et al., 2018), but with metabarcoding this significantly increases the labour required. Since our method can simultaneously sequence hundreds or thousands of species-diagnostic markers, it is likely to offer far greater species-specificity (although this comes at a higher sequencing cost compared to metabarcoding). For individual-level root identification, while clearly superior to existing microsatellite-based methods, our method currently requires all individuals to be present in the catalogue. This makes studies of hundreds of individuals across large geographic areas unpractical for now. Nevertheless, the method could still be effectively applied to large areas by spacing smaller plots (like that used here) across the region, and combining or comparing results across plots. An important future advance would come from developing a reliable exclusion probability statistic for this method, such as that used in paternity testing (Cifuentes et al., 2006). This would allow a measure of certainty of root individual identity even when all individuals are not present in the catalogue. This is not straightforward for GBS data however: exclusion probabilities require knowledge of mutation rates (Cifuentes et al., 2006), yet GBS loci are expected

to be approximately randomly distributed across the genome, including in both highly conserved genic regions and highly variable intergenic regions. Future work on species with ample genomic resources, would allow these regions to be differentiated, and may help to develop an exclusion probability method that is generally applicable.

Technological advancements are opening new fields of study in plant science, particularly in understudied regions like the Caatinga. For example, our method could be combined with techniques such as coarse root distributions derived from e.g., ground penetrating radar (Guo *et al.*, 2013; Almeida *et al.*, 2018) and field sequencing-based plant identification (Parker et al., 2017), to produce highly detailed maps of the root networks of coexisting trees in poorly-studied environments. Our method provides a level of detail which was not previously possible, and has applications across ecology, forestry and agricultural biology.

Acknowledgements

We thank Herica Carvalho, Joabe Almeida, Leide Oliveira and Marcelo Silva for assistance with plant measurements; Italo Coutinho, Bartosz Majcher and Rumbi Chevene for assistance with plant sampling; and Embrapa Semiárido for access to the study site facilities. Access to genetic material from Brazilian plants was authorized by SisGen No. A10E5E1. The work forms part of the joint UK/Brazilian “Nordeste” project funded by the UK Natural Environment Research Council (award numbers NE/N012526/1 and NE/N012550/1) and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP grant number 15/50488-5).

Author contributions

VS, JL, RTP and OGO developed the initial idea, with VS, JL and OGO subsequently designing the experiments; VS supervised the research; OGO created the software package with input from MPD and ASTP; OGO and MPD conducted the data analysis; VS, JL, MSBdM, ATB, LPdQ and RTP conducted the fieldwork; OGO and MPD wrote the first draft of the manuscript and all authors contributed to the final version of the manuscript.

Data availability

Sequence data used in this manuscript has been deposited at the European Nucleotide Archive (accession: tbc). The *RootID* package is available at https://github.com/ogosborne/RootID. All code and additional metadata files required to reproduce the analyses are available at https://github.com/ogosborne/Caatinga_RootID.

References

Agresti, A. 2002. *Categorical Data Analysis*, Second Edition. Hoboken, New Jersey: John Wiley & Sons.

Almeida ER, Porsani JL, Booth A, Brunello AT, Säkinen T. **2018**. Analysis of GPR field parameters for root mapping in Brazil's caatinga environment. *2018 17th International Conference on Ground Penetrating Radar, GPR 2018*: 1–6.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. **2016**. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* **17**: 81–92.

Barberán A, Mcguire KL, Wolf JA, Jones FA, Wright SJ, Turner BL, Essene A, Hubbell SP, Faircloth BC, Fierer N. **2015**. Relating belowground microbial composition to the taxonomic, phylogenetic, and functional trait distributions of trees in a tropical forest. *Ecology Letters* **18**: 1397–1405.

Bardgett RD, Mommer L, De Vries FT. **2014**. Going underground: root traits as drivers of ecosystem processes. *Trends in Ecology and Evolution* **29**: 692–699.

Benjamini Y, Hochberg Y. **1995**. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289–300.

Braukmann TWA, Ivanova N V., Prosser SWJ, Elbrecht V, Steinke D,

Ratnasingham S, de Waard JR, Sones JE, Zakharov E V., Hebert PDN. **2019**. Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources* **19**: 711–727.

Brunner I, Brodbeck S, Büchler U, Sperisen C. **2001**. Molecular identification of fine roots of trees from the Alps: Reliable and fast DNA extraction and PCR-RFLP analyses of plastid DNA. *Molecular Ecology* **10**: 2079–2087.

Cabal C, De Deurwaerder HPT, Matesanz S. **2021**. Field methods to study the spatial root density distribution of individual plants. *Plant and Soil* **462**: 25–43.

Cerca J, Maurstad MF, Rochette NC, Rivera-Colón AG, Rayamajhi N, Catchen JM, Struck TH. 2021. Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. *Methods Ecol Evol* **12**: 805–817.

Cifuentes LO, Martínez EH, Acuña MP, Jonquera HG. **2006**. Probability of exclusion in paternity testing: Time to reassess. *Journal of Forensic Sciences* **51**: 349–350.

Crotti M, Barratt CD, Loader SP, Gower DJ, Streicher JW. 2019. Causes and analytical impacts of missing data in RADseq phylogenetics: Insights from an African frog (Afrixalus). *Zool. Scr*. **48**: 157–167

Davey JL, Blaxter MW. **2010**. RADseq: Next-generation population genetics. *Briefings in Functional Genomics* **9**: 416–423.

de Lima Araújo E, de Castro CC, de Albuquerque, UP. 2007. Dynamics of Brazilian Caatinga–a review concerning the plants, environment and people. *Functional Ecosystems and Communities* **1**: *15-28.*

Estaki M, Jiang L, Bokulich NA, McDonald D, González A, Kosciolek T, Martino C, Zhu Q, Birmingham A, Vázquez-Baeza Y, *et al.* 2020. QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols in Bioinformatics* **70**: e100.

Fernandes MF, Cardoso D, Pennington RT, de Queiroz LP. 2022. The Origins and Historical Assembly of the Brazilian Caatinga Seasonally Dry Tropical Forests. *Frontiers*

in Ecology and Evolution **24:** 101.

Fu Z, Epstein B, Kelley JL, Zheng Q, Bergland AO, Castillo Carrillo CI, Jensen AS, Dahan J, Karasev A V, Snyder WE. 2017. Using NextRAD sequencing to infer movement of herbivores among host plants. *PloS one* 12: e0177742.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, *et al.* 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* **40**: 1178–1186.

Grewe F, Huang J-P, Leavitt SD, Lumbsch HT. **2017**. Reference-based RADseq resolves robust relationships among closely related species of lichen-forming fungi using metagenomic DNA. *Scientific Reports* **7**: 9884.

Guo L, Chen J, Cui X, Fan B, Lin H. **2013**. Application of ground penetrating radar for coarse root detection and quantification: a review. *Plant and Soil* **362**: 1–23.

Hamblin MT, Rabbi IY. **2014**. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in Cassava (Manihot esculenta). *Crop Science* **54**: 2603–2608.

Heyland A, Hodin J. **2004**. Heterochronic developmental shift caused by thyroid hormone in larval sand dollars and its implications for phenotypic plasticity and the evolution of nonfeeding development. *Evolution; international journal of organic evolution* **58**: 524–38.

Hiiesalu I, Öpik M, Metsis M, Lilje L, Davison J, Vasar M, Moora M, Zobel M, Wilson SD, Pärtel M. **2012**. Plant species richness belowground: Higher richness and new patterns revealed by next-generation sequencing. *Molecular Ecology* **21**: 2004–2016.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A. **2008.** Implementing a class of permutation tests: The coin package. *Journal of Statistical Software* **28**: 1–23.

IUSS (International Union of Soil Science) Working Group WRB 2015. World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources

Reports No. 106. FAO, Rome.

Jackson RB, Moore LA, Hoffmann WA, Pockman WT, Linder CR. **1999**. Ecosystem rooting depth determined with caves and DNA. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 11387–11392.

Jones FA, Erickson DL, Bernal MA, Bermingham E, Kress WJ, Herre EA, Muller-Landau HC, Turner BL. **2011**. The roots of diversity: Below ground species richness and rooting distributions in a tropical forest revealed by DNA barcodes and inverse modeling. *PLoS ONE* **6**.

Kesanakurti PR, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG, Graham SW, Barrett SCH, Hajibabaei M, Husband BC. **2011**. Spatial patterns of plant diversity below-ground as revealed by DNA barcoding. *Molecular Ecology* **20**: 1289–1302.

Kim S. **2015**. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Communications for Statistical Applications and Methods* **22**: 665–674.

Lamb EG, Winsley T, Piper CL, Freidrich SA, Siciliano SD. **2016**. A high-throughput belowground plant diversity assay using next-generation sequencing of the trnL intron. *Plant and Soil* **404**: 361–372.

Lang C, Dolynska A, Finkeldey R, Polle A. **2010**. Are beech (Fagus sylvatica) roots territorial? *Forest Ecology and Management* **260**: 1212–1217.

Magalhaes IS, Whiting JR, D'Agostino D, Hohenlohe PA, Mahmud M, Bell MA, Skúlason S, MacColl ADC. **2020**. Intercontinental genomic parallelism in multiple three-spined stickleback adaptive radiations. *Nature Ecology and Evolution*.

Matesanz S, Pescador DS, Pías B, Sánchez AM, Chacón-Labella J, Illuminati A, de la Cruz M, López-Angulo J, Marí-Mena N, Vizcaíno A, *et al.* 2019. Estimating belowground plant abundance with DNA metabarcoding. *Molecular Ecology Resources* **19**: 1265–1277.

Mokany K, Raison RJ, Prokushkin AS. **2006**. Critical analysis of root: Shoot ratios in terrestrial biomes. *Global Change Biology* **12**: 84–96.

Mommer L, van Ruijven J, de Caluwe H, Smit-Tiekstra AE, Wagemaker CAM, Joop Ouborg N, Bögemann GM, van der Weerden GM, Berendse F, de Kroon H. **2010**. Unveiling below-ground species abundance in a biodiversity experiment: A test of vertical niche differentiation among grassland species. *Journal of Ecology* **98**: 1117–1127.

Moonlight PW, Banda-R K, Phillips OL, Dexter KG, Pennington RT, Baker TR, C. de Lima H, Fajardo L, González-m R, Linares-Palomino R. 2021. Expanding tropical forest monitoring into Dry Forests: The DRYFLOR protocol for permanent plots**.** *Plants, People, Planet* **3:** 295-300.

Murdoch D, Adler D. **2021**. rgl: 3D Visualization Using OpenGL. R package version 0.108.3.

Norrgard K. (2008). Forensics, DNA fingerprinting, and CODIS. *Nature Education* **1**:35

Ostle NJ, Smith P, Fisher R, Ian Woodward F, Fisher JB, Smith JU, Galbraith D, Levy P, Meir P, McNamara NP, *et al.* 2009. Integrating plant-soil interactions into global carbon cycle models. *Journal of Ecology* **97**: 851–863.

Papadopulos AST, Igea J, Dunning LT, Osborne OG, Quan X, Pellicer J, Turnbull C, Hutton I, Baker WJ, Butlin RK, *et al.* 2019. Ecological speciation in sympatric palms: 3. Genetic map reveals genomic islands underlying species divergence in Howea. *Evolution* **73**.

Parker J, Helmstetter AJ, Devey D. et al. 2017. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Scientific Reports* **7**: 8345

Paris JR, Stevens JR, Catchen JM. **2017**. Lost in parameter space: a road map for stacks. *Methods in Ecology and Evolution* **8**: 1360–1373.

Pennington RT, Prado DE, and Pendry CA. 2000. Neotropical seasonally dry forests and Pleistocene vegetation changes. *Journal of Biogeography* **27:** 261–273.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. **2012**. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **7**.

Poorter H, Niklas KJ, Reich PB, Oleksyn J, Poot P, Mommer L. **2012**. Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control. *New Phytologist* **193**: 30–50.

Quesada CA, Lloyd J, Anderson LO, Fyllas NM, Schwarz M, Czimczik CI. 2011. Soils of Amazonia with particular reference to the RAINFOR sites. *Biogeosciences*, **8:** 1415-1440.

Rivera-Colón AG, Rochette NC, Catchen JM. **2021**. Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Molecular Ecology Resources* **21**: 363–378.

Rochette NC, Rivera-Colón AG, Catchen JM. **2019**. Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*: 4737–4754.

Saari SK, Campbell CD, Russell J, Alexander IJ, Anderson IC. **2005**. Pine microsatellite markers allow roots and ectomycorrhizas to be linked to individual trees. *New Phytologist* **165**: 295–304.

Sampaio EVSB, Silva, G. 2016. Biomass equation for Brazilian Semiarid Caatinga plants. *Revista Ciência Agronômica* **47:** 32–40.

Schliep KP. **2011**. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**: 592–593.

Torello-Raventos M, Feldpausch TR, Veenendaal E, Schrodt F, Saiz G, Domingues TF, Djagbletey G, Ford A, Kemp J, Marimon BS *et al.* 2013. On the delineation of tropical vegetation types with an emphasis on forest/savanna transitions. *Plant Ecology & Diversity* **6:** 101-137.

T**umber-Dávila S.J., Schenk H.J., Du E. and Jackson R.B. (2022)** Plant sizes and shapes above and belowground and their interactions with climate. *New Phytologist* **235**: 1032-1056.

Xin Z, Chen J. **2012**. A high throughput DNA extraction method with high yield and quality. *Plant Methods* **8**: 26.

Zane L, Bargelloni L, Patarnello T. **2002**. Strategies for microsatellite isolation: A review. *Molecular Ecology* **11**: 1–16.

Zhang GK, Chain FJJ, Abbott CL, Cristescu ME. **2018**. Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications* **11**: 1901–1914.

Figures

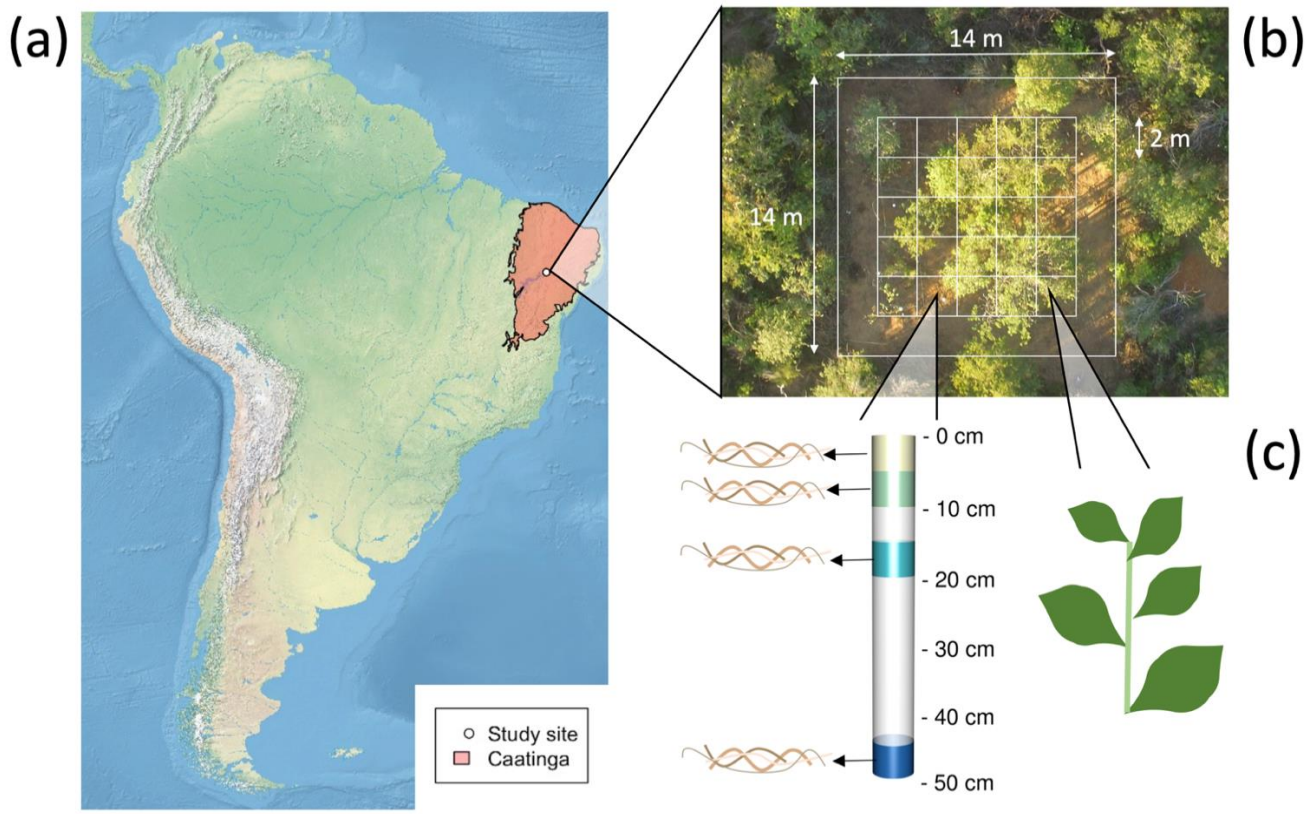

Figure 1. A schematic showing our sampling strategy. A map shows the location of the experimental plot within the Caatinga region in South America (a). The sampling design is superimposed onto an aerial photograph of the plot (b): the 10 × 10 m central section of the plot is divided into 2 × 2 m subplots and a soil core is taken from the centre of each subplot (represented as a cylinder in panel c). Roots are sampled from four different depth ranges in each soil core (coloured sections in panel c) and leaves are sampled from all trees and shrubs within the 14 × 14 m plot. The background map image was created from the Natural Earth 2 dataset (naturalearthdata.com) which is free to use without restriction, and all other images are the authors' own work.

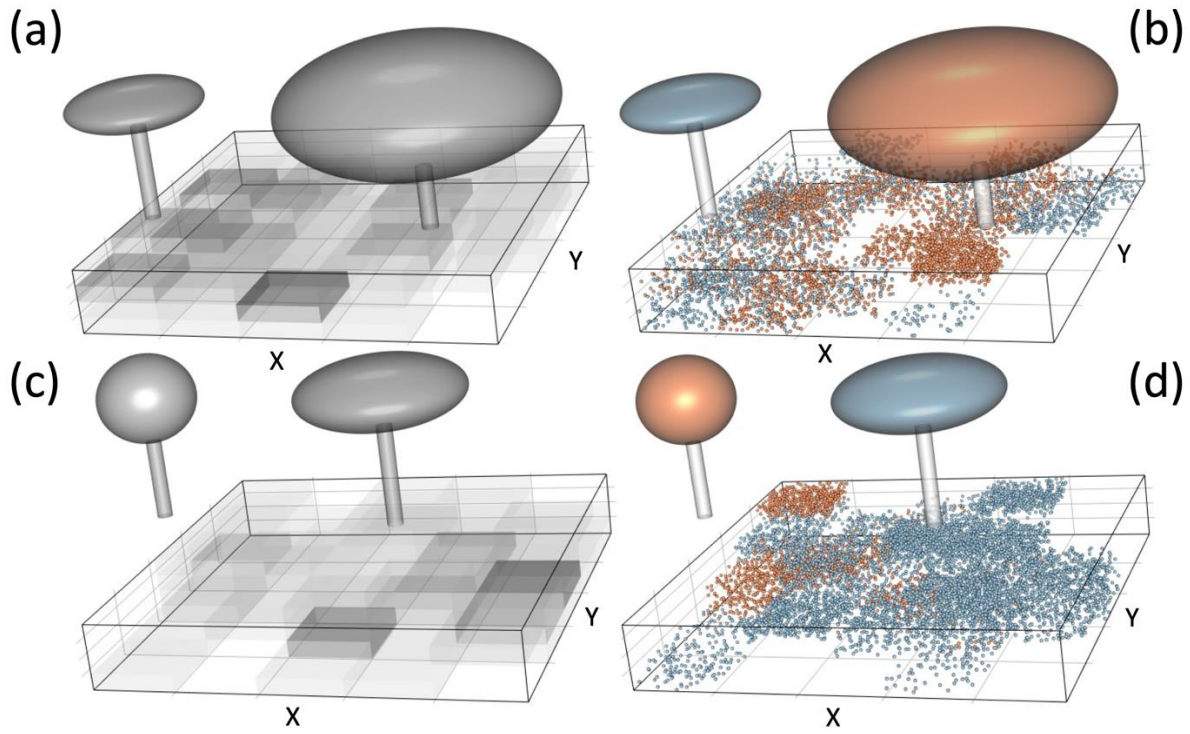

Figure 2. The estimated root distribution of two of the study species: *Commiphora leptophloeos* (panel a) and *Cenostigma microphyllum* (panel c) based on species-diagnostic markers and the estimated root distribution of the individuals of these species based on individual-diagnostic haplotypes (*Commiphora leptophloeos*: panel b; *Cenostigma microphyllum*: panel d). Panels (a) and (c) show the number of species-diagnostic marker reads for each species scaled by the maximum number found in any subplot. To show more easily rooting depth, we represented these as transparent cuboids, where darker colours indicate more species-diagnostic markers. Panels (b) and (d) show the proportion of individual-diagnostic haplotypes of each individual, scaled by the maximum proportion found for any individual. To show multiple individuals within the plot, we represented these as randomly distributed points within each subplot, where higher point density indicates higher relative abundance. Points are coloured by the tree they are associated with. Each panel shows a map of all 2 × 2 × 0.05 m subplots with each subplot represented as a cuboid. Tree models show the location, canopy area, tree height and crown base height of the trees. Axis labels show the axis

identifiers (see Table S2). Gridlines in the horizontal plane show the horizontal extent of each subplot and vertical gridlines show the four sampling depth levels: 0–5 cm, 5–10 cm, 15–20 cm and 45–50 cm, from top to bottom. Root sampling depths are not to scale but the horizontal root axes and the trees are.

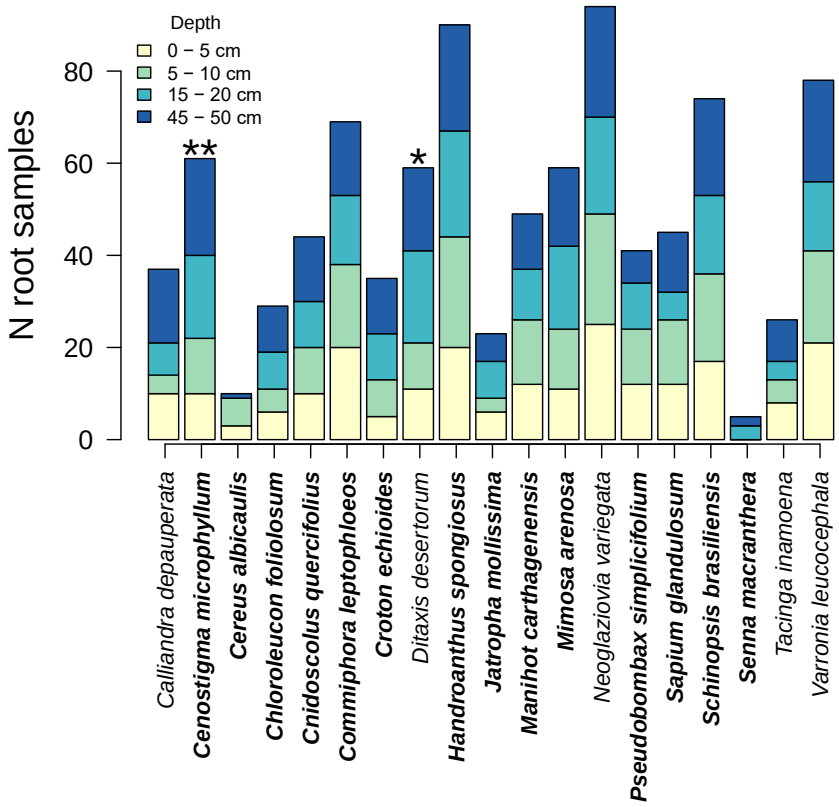

Figure 3. Depth distribution of each species. Each bar is divided into four sections, showing the number of root samples each species was detected in at each of the four sampling depths (0–5 cm, 5–10 cm, 15–20 cm and 45–50 cm). Stars above the bars indicate that species detection or non-detection was significantly associated with sampling depth following correction for multiple testing (linear-by-linear association tests; one star < 0.05, two stars < 0.01). Names of tree/shrub species are shown in bold-italic and subshrub/herb species are shown in italic.

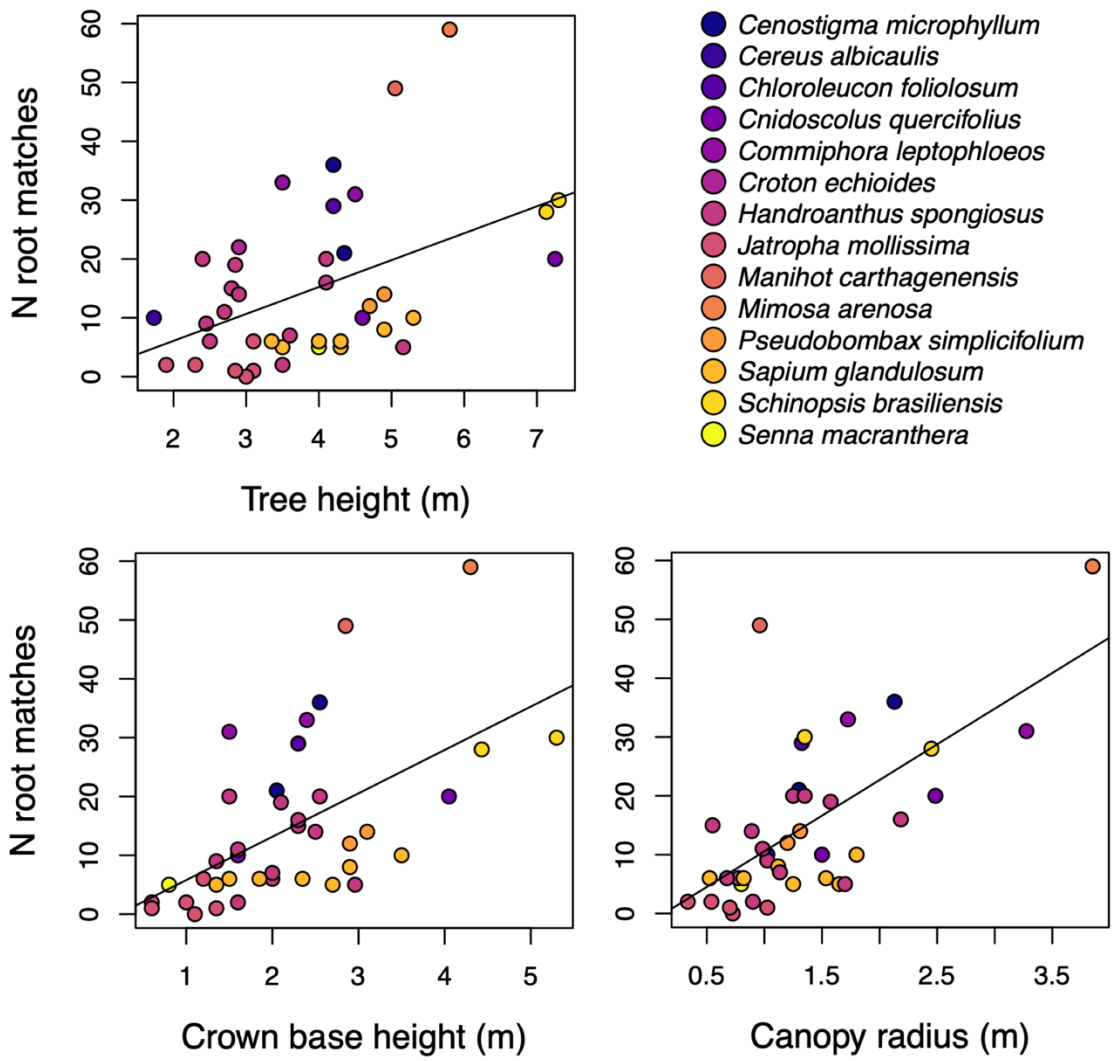

Figure 4. The relationship between aboveground measurements (tree height, crown base height and canopy radius) and the number of root samples each individual was detected in. Each point represents an individual tree, and points are coloured by species. The line shows the linear regression.