nature communications

Article

https://doi.org/10.1038/s41467-024-54940-5

Increased diversity of beneficial rhizobia enhances faba bean growth

Received: 5 May 2024

Accepted: 22 November 2024

Published online: 12 December 2024

Check for updates

Marcela Mendoza-Suárez[1,2] ✉, Turgut Yigit Akyol[1,2], Marcin Nadzieja[1] & Stig U. Andersen[1] ✉

Legume-rhizobium symbiosis provides a sustainable nitrogen source for agriculture. Nitrogen fixation efficiency depends on both legume and rhizobium genotypes, but the implications of their interactions for plant performance in environments with many competing rhizobium strains remain unclear. Here, we let 399 *Rhizobium leguminosarum* complex sv. *viciae* strains compete for nodulation of 212 faba bean genotypes. We find that the strains can be categorised by their nodule occupancy profiles into groups that show distinct competitive interactions and plant growth-promoting effects. Further, we show that the diversity of strains occupying root nodules affects plant growth and is under plant genetic control. These insights provide a basis for re-designing rhizobium inoculation and plant breeding strategies to enhance symbiotic nitrogen fixation in agriculture.

Faba bean (*Vicia faba*) is a globally adapted protein crop with the highest yield potential of all grain legumes and is characterised by efficient symbiotic nitrogen fixation[1,2]. With recent advances in genetic and genomic resources, including a reference genome and genetic characterisation of germplasm collections[3,4], faba bean is now amenable to genetic studies of complex traits. Nitrogen fixation is carried out by symbiotic rhizobia in root nodules, where faba bean and pea (*Pisum sativum*) are mainly nodulated by *Rhizobium leguminosarum* complex (*Rlc*) sv. *viciae*[5,6] while other *R. leguminosarum* complex symbiovars nodulate additional important crop legumes, including white clover (*Trifolium repens*) and common bean (*Phaseolus vulgaris*). The compatibility with the legume host, defining the symbiovar, is determined by symbiosis genes, which can be transferred within the *R. leguminosarum* species complex[5–9]. Here, we will refer to *Rlc* sv. *viciae* strains using the abbreviation *Rlcv*.

In soil, rhizobia exist in complex communities, with potentially hundreds of compatible strains available within the soil volume accessible to the legume host[8,9]. Legume-rhizobium symbiotic nitrogen fixation, takes place in an environment where many rhizobium strains are competing for occupying legume root nodules to gain access to photosynthates, and where the legume attempts to select the best symbiotic partner to maximise nitrogen fixation[10]. This complex situation is difficult to recapitulate and examine in controlled experiments, but multi-strain inoculation has been used as an approximation[11,12]. Some of these studies compared the effects of single- versus multi-strain inoculum, reaching different conclusions on their relative efficacy[13–15]. Others focused on studying the relationships of host and microbial fitness[12,16–18].

Even without inoculation, faba bean has shown consistently high levels of nitrogen fixation in British soils, characterised by a high abundance of *Rlcv* strains[19], whereas a number of studies in African soils have shown significant effects of inoculation, likely because limited availability of *Rlcv* strains offered the inoculants a competitive advantage[20]. *Rhizobium* competitiveness for nodulation and nitrogen fixation efficiency are independent traits, and, with respect to inoculant development, the objective is to identify strains that are both competitive and efficient[21]. A key challenge is that ascribing a specific growth-promoting effect to individual rhizobium strains in a complex mixture is difficult because the effects of single strains will likely be small. Here, we present large-scale data from 399 *Rlcv* strains competing for nodulation of 212 faba bean genotypes, along with new analysis approaches, to address these challenges and link *Rlcv* growth-promoting effects through community diversity to plant genetics.

[1]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark. [2]These authors contributed equally: Marcela Mendoza-Suárez, Turgut Yigit Akyol. ✉e-mail: marcela@mbg.au.dk; sua@mbg.au.dk

Results

A diverse *Rlcv* library for competition studies

To construct a comprehensive collection of diverse *Rlcv* strains, we collected soil samples from ten locations across six European countries (Fig. 1a) from October 2019 until February 2020. For each sample, we recorded GPS coordinates, pH, nutrient availability, organic material, soil type and texture, *Rlcv* concentration using most probable number (MPN) analysis[22], previous use of the land, and previous legume cropping (Fig. 1b, and Supplementary Data 2). MPN was performed as soon as the soil was collected to ensure the viability of the *Rlcv* cells. We observed significant geographical variation in *Rlcv* concentrations. Spanish soils, for instance, exhibited fewer than 20 cells/g soil, while other soils contained over 58,000 cells/g (Supplementary Data 2).

We subsequently used each soil in combination with six diverse faba bean genotypes (Hedin/2, Melodie, Alameda, ILB938/2, VF172/3cv, and Giza402Gö) to trap *Rlcv* strains (Fig. 1c and Supplementary Fig. 7a, b). Approximately 10,000 nodules were harvested and surface-sterilized, and *Rlcv* strains were isolated (Supplementary Fig. 7c). We then performed enterobacterial repetitive intergenic consensus (ERIC) PCRs[23] to generate a DNA fingerprint profile for each isolated strain, and besides the *Rlcv* concentrations, we also observed diversity differences across soils. For example, the SEJET soil from Denmark exhibited high *Rlcv* concentration but low diversity, while the IFAPA soil from Spain showed low concentration but high diversity (Supplementary Data 2 and Supplementary Fig. 8a, b). We selected 452 unique strains based on their ERIC PCR fingerprinting profiling and conjugated them with a new version of the Plasmid-ID[24]. This version contains a gentamicin antibiotic cassette in the backbone vector[25] (Fig. 1e and Supplementary Fig. 9). The final tagged library comprised 399 strains. Initially, soil samples from ten sites were used to trap *Rlcv* isolates. However, the final tagged strain library did not equally represent each soil since the strain diversity, based on ERIC PCR, was very low in some soils (Supplementary Fig. 8a, b) and in a few cases, the high-throughput conjugation was not successful.

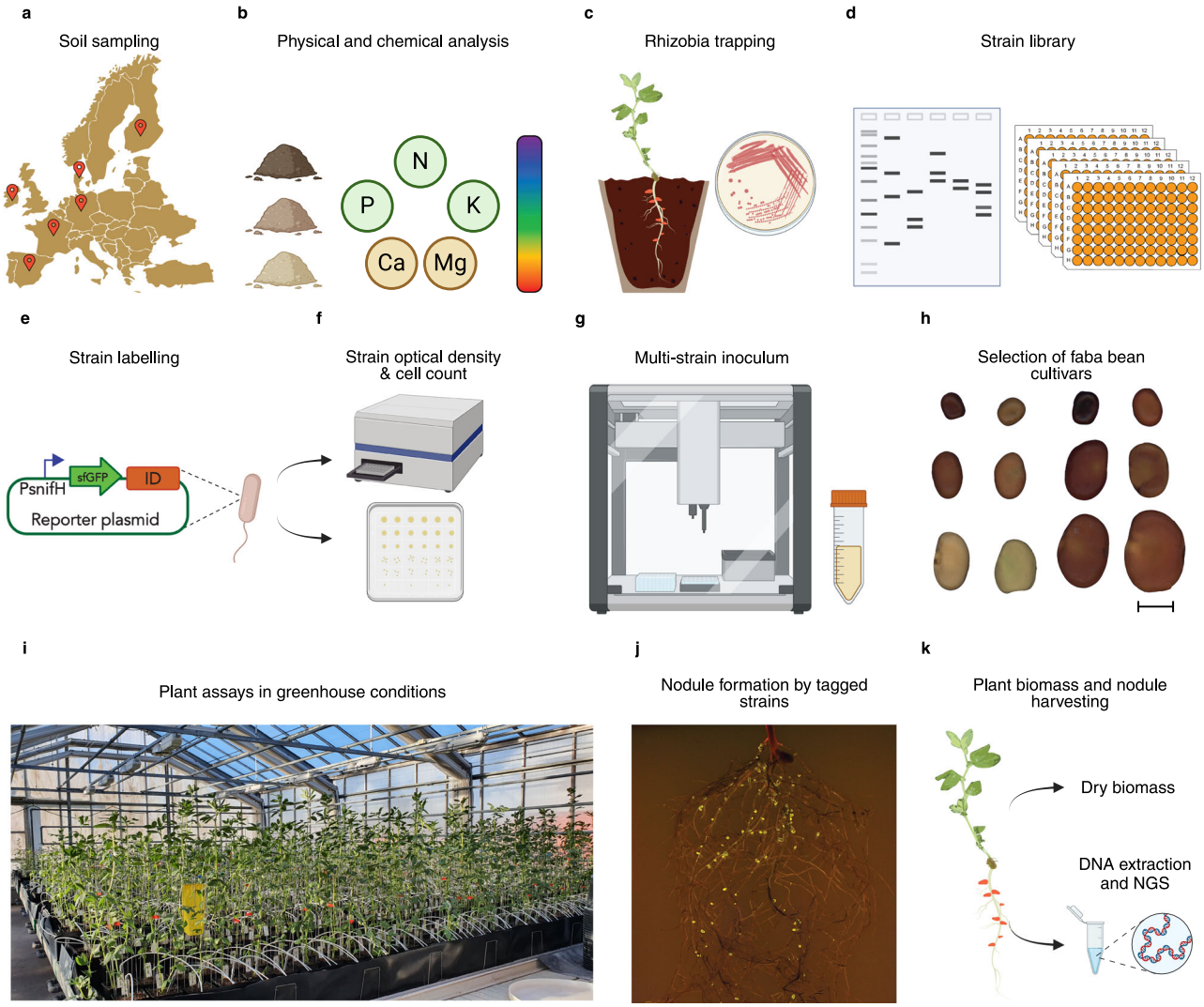

Fig. 1 | Analysing rhizobia competitiveness and efficiency. a Soil sample collection sites. **b** Physical and chemical soil analysis. **c** Trapping *Rlcv* isolates with faba bean plants by plating bacteria harvested from surface-sterilised nodules. **d** ERIC-PCR fingerprinting of *Rlcv* strains to select unique strains. **e** Labelling of selected strains with the Plasmid-ID system. **f** Verification of the strain library by bacterial cell counting using optical density (OD) and serial dilutions on plates. **g** We use the OD measurements to program the OT-2 pipetting robot to transfer the volume needed to create the final inoculant with a comparable number of cells from each strain. **h** We used our multi-strain inoculum (+R) to inoculate 212 faba bean genotypes from the ProFaba diversity panel. **i** Three biological replicates of each plant genotype and each condition were grown under greenhouse conditions in eight batches. **j** Roots were exposed to a blue-light transilluminator to verify the presence of Plasmid-ID-labelled strains. **k** Plant shoot dry biomass was quantified and root nodules were surface-sterilized, pooled and DNA was extracted for Illumina sequencing (NGS). Created in BioRender. Mendoza, M. (2023) BioRender.com/j32h564.

Competitive characteristics of the *Rlcv* strains

We assessed the competitive characteristics of the *Rlcv* strains across the 212 faba bean genotypes from the ProFaba diversity panel, which captures global faba bean diversity and includes inbred lines derived from commercial cultivars (Fig. 1h)[4]. We prepared a multi-strain inoculum with 399 tagged *Rlcv* strains at a final density of 2×10^4 cell/ml (Fig. 1f, g). Each pot with the inoculation treatment was inoculated with 1 ml of the multi-strain inoculum.

In addition, we grew the faba bean genotypes with full chemical fertilisation as a positive control and under nitrogen starvation as a negative control. We cultivated the plants in three biological replicates of each treatment, dividing the experiment into eight batches using a randomized setup under greenhouse conditions (Supplementary Data Table 1). To ensure consistency in the inoculation across all batches, the multi-strain inoculum was prepared each time from −80 °C stocks and we normalized each strain by their optical density (OD) (Fig. 1g). We then sequenced the inoculum from each batch and found that the vast majority of the isolates had very similar average abundances across the eight inocula (Supplementary Fig. 1a–c). Only a few of the isolates were over- or poorly represented in an inoculum and the pairwise community comparisons showed only two significant differences (inoculum 3 vs 8 and 4 vs 8, Supplementary Fig. 1d). Nevertheless, to prevent any biases that could result from variations in isolate abundances across batches, we normalised the isolate counts in the nodule samples by dividing them by the counts observed in the inocula.

After 60 days, we harvested the plant shoot and recorded its dry biomass (Fig. 1i). Simultaneously, the rhizobia-inoculated roots were inspected using a blue-light transilluminator to verify that they were colonized by Plasmid-ID-labelled strains (Fig. 1j and Supplementary Fig. 10). We harvested all root nodules from each individual plant, surface-sterilized them, and pooled them. We then extracted DNA from the pooled nodules and carried out multiplexed Illumina sequencing of the Plasmid-ID region, to determine the relative occupancy of each strain (Fig. 1k). Sequencing pooled samples means that the relative occupancy does not reflect the number of nodules in which each strain occurs. Instead, the relative occupancy of a strain reflects the fraction its DNA constitutes of the total pool of *Rlcv* DNA in each sample. A few large nodules can thus contribute as much to the occupancy as many smaller ones.

Out of the initial 399 tagged strains in the inoculum, we identified 397 in the sequencing data across all plant samples. The read counts per sequencing library in the plant samples before the inoculum normalisation ranged from 25 to 333,566 with a median count of 113,866. We excluded five samples with read counts below 2000, leaving 603 samples. The nodule occupancies appeared highly strain-dependent, with some strains showing consistently high occupancies across most faba bean genotypes (Fig. 2a). We, therefore, classified the strains into four groups according to their nodule occupancy profiles based on occurrence (the number of plants in which the strain was detected) and abundance (the average relative abundance of each strain) (Fig. 2b, Supplementary Fig. 2). We named the four groups: Dominants (high occurrence/high abundance), Specialists (low occurrence/high abundance), Generalists (high occurrence/low abundance), and Transients (low occurrence/low abundance) (Fig. 2b). The soil from UG (University of Göttingen, Germany), SJ and ND (Sejet and Nordic Seed, Denmark, respectively) contributed most of the competitive strains, whereas the remaining soils (Fig. 2 Others, Supplementary Table 1) mainly contributed strains that were unsuccessful in nodule colonization (Fig. 2c). Out of 86 Dominant strains, 80 originated from UG soil. Similarly, a large proportion of Specialist strains (54 out of 73) were also from UG soil. In contrast, SJ and ND strains were mostly placed in the Transient and Generalist groups, with 22 and 28 out of a total of 57 Generalist strains originating from SJ and ND soils, respectively. Further, more than 50% of the isolates from either SJ or ND were classified as Transients (Supplementary Table 1). The inter-group differences were also pronounced with respect to niche breadth[26], a metric that describes the uniformity of the distribution of the strains across environments (plant samples) (Fig. 2d).

Rlcv community dynamics

Next, we investigated if the groups also showed distinct community dynamics. Leaving out the Transients, we investigated the interactions among the strains within the remaining three groups based on their co-occurrence, mutual exclusion, correlation, and host-dependency (Fig. 3, Supplementary Fig. 3). The Dominants generally occurred together, did not show any mutual exclusion and their abundances were not correlated (Fig. 3a–c). The Generalists showed both co-occurrence and strongly correlated abundance profiles (Fig. 3e, g) without mutual exclusion (Fig. 3f). The Specialists were unique in showing no co-occurrence, but instead a large number of mutual exclusions among isolate pairs (Fig. 3I, k). We also applied dissimilarity-overlap curve (DOC) analyses to the groups[27]. A negative DOC slope between the number of overlapping taxa (rhizobium strains) and the dissimilarity of the subject pairs (faba bean genotypes) indicates that the strains interact similarly in distinct individuals when they occur together[28]. The Generalists displayed this profile, indicating a host-independent colonisation pattern (Fig. 3h), whereas the Dominants and Specialists showed host-dependent patterns where the dissimilarity did not decrease with increasing strain overlaps (Fig. 3d, l). Examining the between-group relationships, we found a strong negative correlation between Dominants and Specialists, suggesting that they were competing directly (Fig. 3m). In contrast, the Generalists interacted little with the other groups, just as they showed no interaction with plant genotype (Fig. 3h, n–p).

Rlcv groups have distinct effects on plant growth

Since the groups showed pronounced differences in their nodule occupancy and interaction characteristics, we investigated if they also differed in their effect on plant growth. We used a linear mixed model to assess the effect of each strain, controlling for the plant genotype, the batch effect, and the community structure of the rhizobium strains (Supplementary Fig. 4). As expected, with many strains present in the inoculum, the effect of individual strains was small, and we identified only two strains with a significant influence on plant growth (Fig. 4a). At the group level, however, we saw significant differences ($P < 0.05$, Tukey's HSD test) (Fig. 4b). Specialists were most beneficial, closely followed by Dominants, whereas Generalists had a more negative effect on plant growth (Fig. 4b). Furthermore, Generalists showed a negative correlation between niche breadth and plant growth effect ($r = -0.39$, $P = 0.004$), in contrast to Dominants and Specialists (Fig. 4c).

Rlcv diversity is under plant genetic control

Overall, the different groups had distinct effects on plant growth when modelling the effect of one strain at a time (Fig. 4b). To be able to link *Rlcv* traits to plant genetics, we then asked if we could explain variation in plant growth using parameters that summarise the *Rlcv* nodule communities of individual plants. Because of the negative effect of the Generalists, we first tested the impact of the cumulative abundances of the different classes. However, the group cumulative abundances did not explain a significant proportion of plant growth variance, they did not improve the prediction of plant growth and Generalist cumulative abundance was difficult to predict based on plant genetic data (Supplementary Fig. 5). Since the groups also showed different community characteristics, we next examined a number of different community summary metrics (alpha and beta diversity) for their ability to improve prediction of plant growth when considered together with plant genotype information. Shannon's diversity and evenness significantly improved the prediction accuracy (Fig. 4d). In contrast to group

https://doi.org/10.1038/s41467-024-54940-5

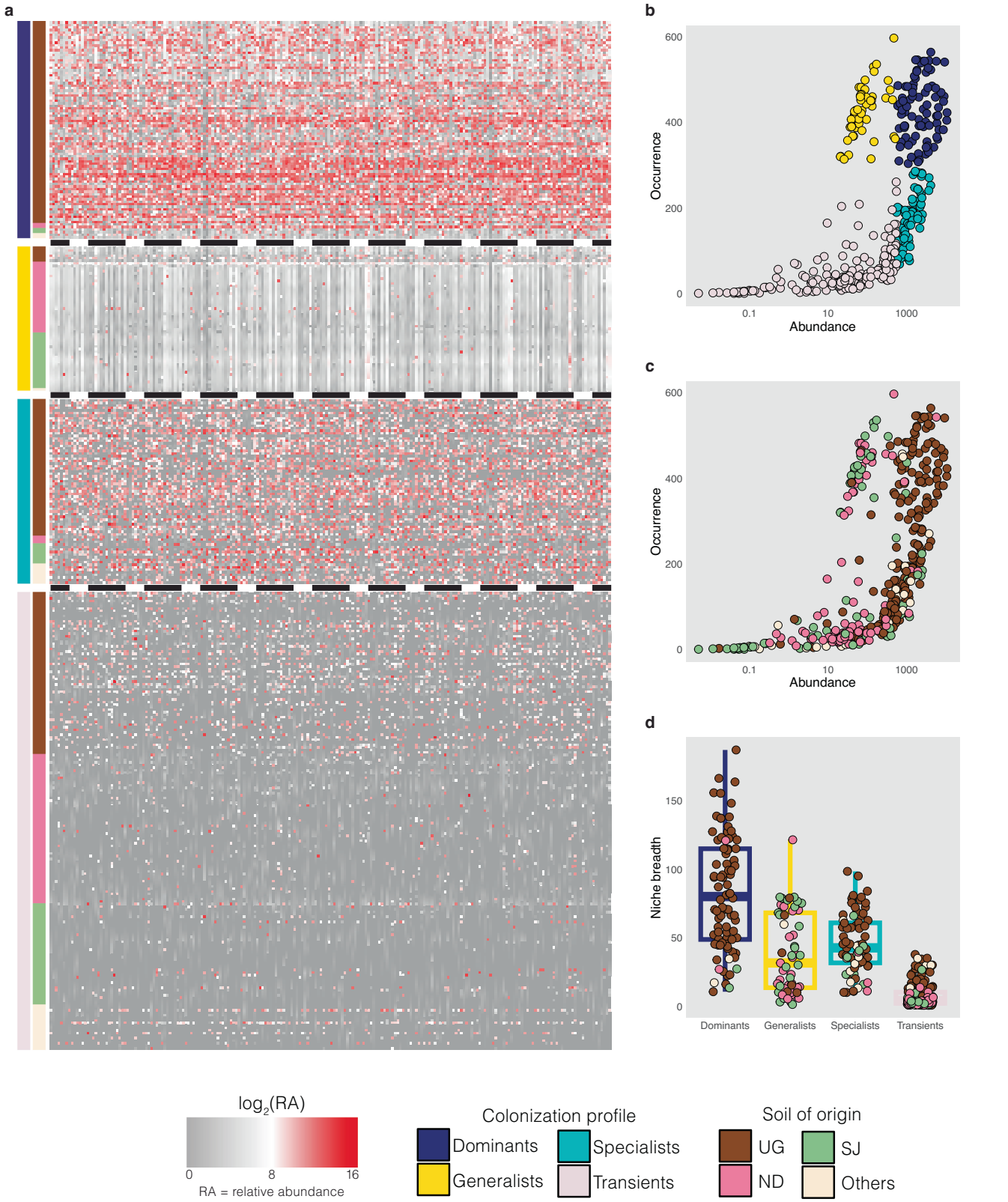

Fig. 2 | *Rhizobium* nodule occupancy. a Heatmap showing the $\log_2$-transformed average relative abundances of the *Rhizobium* strains (*y*-axis) in 212 faba bean genotypes (*x*-axis). Vertical bars at the left denote their associated group and their soil origin. **b**, **c** Abundance-occurrence plots of *Rhizobium* strains. Average relative abundance of each strain is shown on the $\log_2$-scaled *x*-axis. Occurrence indicates the number of samples in which an isolate was detected. Strains are coloured by colonisation groups in (**b**) and by soil origin in (**c**). **d** Boxplot depicting the niche breadth of the strains by colonisation group. $n = 86$, 57, 73 and 181 (from left to right). Boxplot indicates median (middle line), 25th, 75th percentile (box) and 5th and 95th percentile (whiskers) as well as all data points. Strains are coloured by soil origin. UG: University of Göttingen, Germany; SJ: Sejet, Denmark; ND: Nordic Seed, Denmark; Others: the remaining soils. See Supplementary Data 2 for detailed information of soils. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-54940-5

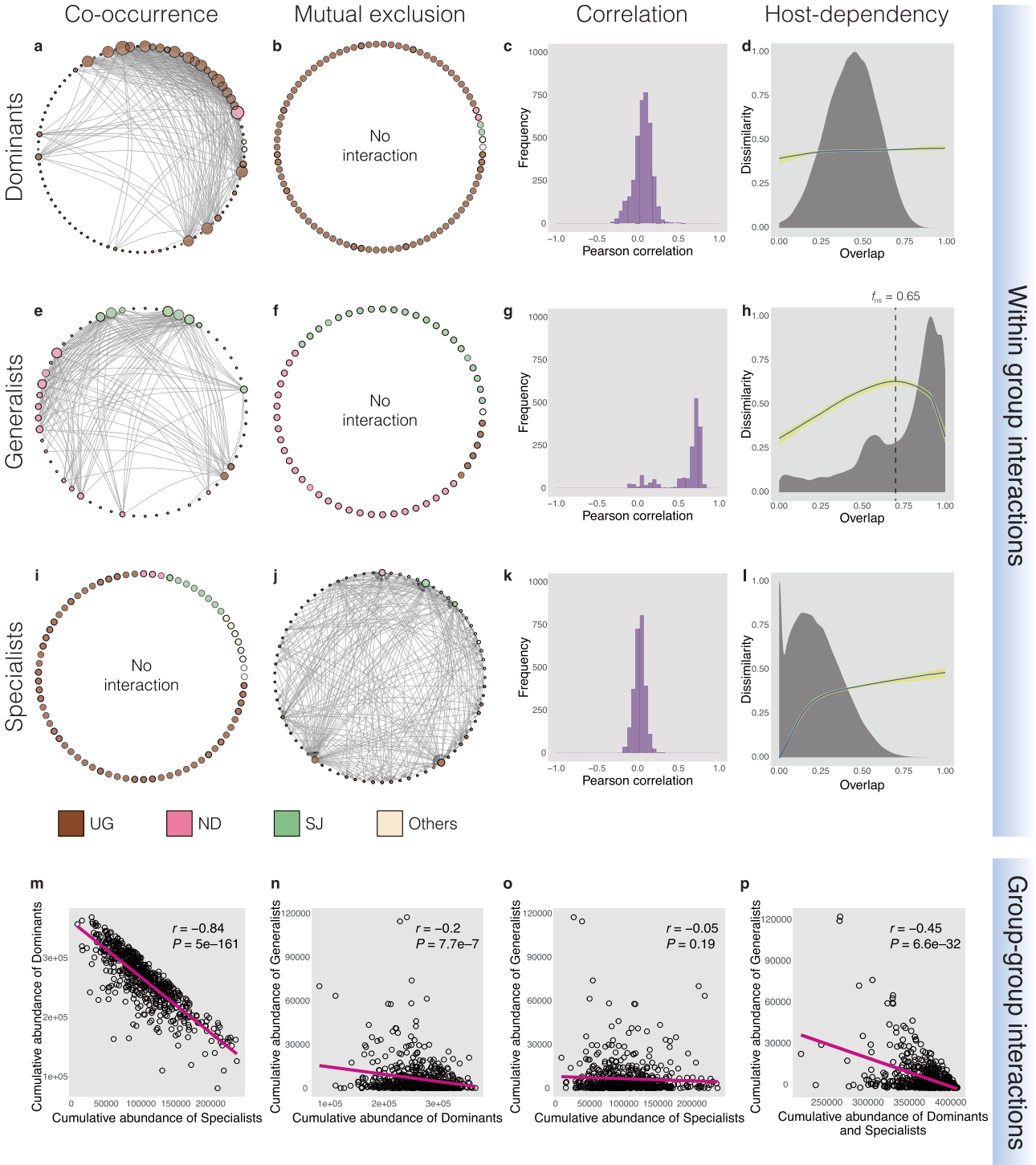

Fig. 3 | Rhizobia interactions. a, **e**, **i** Co-occurrence plots show the presence-presence interactions among the rhizobium strains (circles). A grey line between a pair of isolates denotes that they co-occurred in many samples while one occurred without the other in few samples (Supplementary Fig. 3). **b**, **f**, **j** Mutual exclusion plots show the presence-absence interactions among the rhizobium strains. Competitive superiority is indicated by arrows pointing away from the circle (Supplementary Fig. 3). Isolates (circles) are coloured with respect to their soil origin (UG: University of Göttingen, Germany; SJ: Sejet, Denmark; ND: Nordic Seed, Denmark; Others: the remaining soils.) and sized according to their number of interactions. **c**, **g**, **k** Histograms for Pearson correlation coefficient (r) between all the pairs of strains. Correlations are calculated based on the relative abundances of the isolates. **d**, **h**, **l** Dissimilarity overlap curves (DOCs) (blue lines) calculated using the robust LOWESS (locally weighted scatterplot smoothing) method. Confidence intervals for DOCs (yellow shaded areas) represent 2.5th and 97.5th percentiles of the curves calculated from 100 bootstraps. The vertical dashed line indicates the point where a negative DOC is first observed. Dissimilarity is based on Jensen–Shannon distance. The density of sample pair strain co-occurrence is shown in grey. **m**, **n**, **o**, **p** Correlation plots of the cumulative abundances of the indicated group combinations. Each scatter plot includes Pearson correlation coefficient (*r*) and *P*-value. All tests are two-sided. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-54940-5

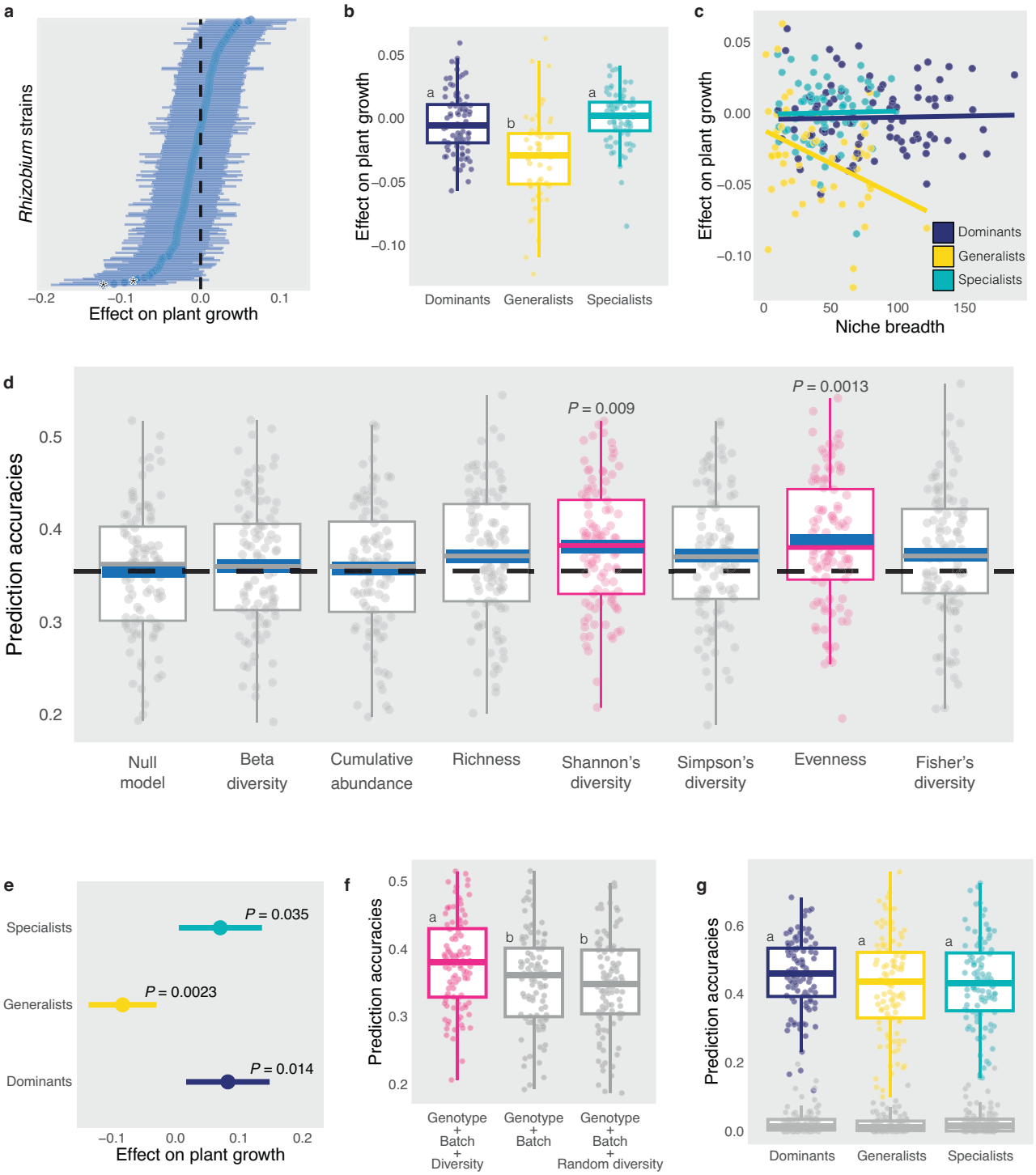

Fig. 4 | Rhizobia effects on plant growth. a Effect of strain abundance on plant growth estimated using linear mixed model (LMM) analysis. Estimates are shown as dots with confidence intervals extending from 2.5th to 97.5th percentile. Asterisks denote statistical significance (FDR < 0.05). **b** Estimates shown in (**a**) binned by colonisation group. Significance was determined via ANOVA; letters correspond to a Tukey post hoc test. **c** Strain effect on plant growth plotted versus its niche breadth. Strains are coloured by colonisation group including the best linear regression fit for each group. **d** Random forest (RF) prediction of plant growth based on plant genotype, batch, and *Rhizobium* community parameters. Significance was determined via a two-sided *t*-test. Blue horizontal lines indicate means. The horizontal dashed line indicates the mean null model accuracy. **e** Effect of diversity on plant growth estimated using LMM analysis. Estimates are shown as dots with confidence intervals extending from 2.5th to 97.5th percentile. Significance was determined via ANOVA. **f** RF prediction of plant growth based on plant genotype, batch and diversity. Significance was determined using ANOVA; letters correspond to a Tukey post hoc test (*P*-values: Genotype + Batch + Diversity vs Genotype + Batch = 0.02; Genotype + Batch + Diversity vs Genotype + Batch + Random diversity = 0.003; Genotype + Batch vs Genotype + Batch + Random diversity = 0.77). **g** RF prediction of diversity based on plant genotype. Significance was determined via ANOVA; letters correspond to a Tukey post hoc test (*P*-values: Dominants vs Generalists = 0.16; Dominants vs Specialists = 0.47; Generalists vs Specialists = 0.79). All boxplots indicate median (middle line), 25th, 75th percentile (box) and 5th and 95th percentile (whiskers) as well as all data points. In **a**, **b**, **e** $n = 601$ samples from 8 batches. Source data are provided as a Source Data file. In **d**, **f**, **g** Prediction accuracies (R^2 values) from 100 RF analyses with random test-training splits are shown.

cumulative abundance, Shannon's diversity explained a significant proportion of variation in plant growth (Fig. 4e). Increased Shannon's diversity is associated with improved plant growth for Dominants and Specialists. Moreover, adding Shannon's diversity information for all three groups improved prediction of plant growth (Fig. 4f) and Shannon's diversity could be predicted for all three groups based on plant genetic data using a combination of genome-wide association (GWA) analysis and random forest machine learning (Fig. 4g). The quality of the Shannon's diversity prediction was reflected in the marker consistency, with many markers selected in more than 60 out of 100 cross-validation iterations (Supplementary Fig. 6a–c). This contrasted with the prediction of Generalist cumulative relative abundance, where marker selection consistency was low (Supplementary Fig. 6e).

Discussion

Using six faba bean genotypes to trap compatible rhizobia from different soils, we found pronounced cross-compatibility, with a large proportion of the strains successfully colonising all 212 inoculated faba bean genotypes to some degree (Fig. 2a). This is reminiscent of a study of the interactions between white clover and *R. leguminosarum* bv. *trifolii* (*Rlt*), where all tested pairwise clover-rhizobia combinations were compatible[29]. In that case, *Rlt* genetic variation was not found to contribute to variation in plant growth, which was tentatively attributed to only using strains from large and healthy-looking nodules from the initial trapping experiment[7]. In the current study, we used strains from nodules of all appearances to capture a wider range of diversity, including potentially less beneficial strains.

The ability to sample a large set of diverse rhizobia, and to inoculate and subsequently quickly quantify the nodule occupancy of hundreds of strains, was key to generating sufficient data for identifying the distinct *Rlcv* groups and for robust statistical analysis of inter-group differences. This was made possible by expanding the Plasmid-ID system[24] to a collection of 475 tagged plasmids easily transformed into *R. leguminosarum*, whereas other studies have been limited by the requirement to characterise nodule occupancy using whole-genome sequencing or to use single-strain inoculation and pairwise testing[18,29].

Our large dataset allowed us to regress plant growth on the nodule occupancy of each strain in a mixture, which has been suggested as the potentially best approach to compare strain benefits in a way the reflects complex natural or agricultural environments[30]. Since most strain effect estimates were small and not statistically significant, the results of the regression do not allow confident selection of specific strains for inoculant use and would have been hard to interpret without the additional information about the *Rlcv* group membership (Fig. 4a, b). Indeed, our results indicate that the highly abundant Dominants and Specialists have more positive effects on plant growth than the frequently occurring but less abundant Generalists (Figs. 2b and 4b). This is consistent with legume sanctioning of rhizobia based on nitrogen output only taking effect after nodules have formed and been colonised[31], suggesting that Generalist nodules may generally be smaller than those of Dominants and Specialists. Because we sequenced pooled nodule samples, we could not determine if such a correlation exists, but it would be an interesting topic for future studies. The groups were also key to linking *Rlcv* symbiotic performance to plant genetics since we could identify a community summary statistic, Shannon's diversity, which was associated with better plant growth, improved prediction of plant growth and could be predicted based on plant genetic data (Fig. 4e–g). Because of the key role of the groups in data analysis, we repeated all analyses with a different cut-off for group membership, which changed group membership for 39 strains, including the most frequently occurring strain. The analysis results were robust to the changed threshold, including consistent selection of markers in genomic prediction of Dominant and Specialist *Rlcv* nodule diversity (see Supplementary Note).

The identification of functionally distinct *Rlcv* groups, and the link between their nodule community profiles and plant genetics, present new opportunities for inoculation design and plant breeding strategies. With the large numbers of strains in each *Rlcv* group, it may now be possible to develop genetic markers to rapidly differentiate between groups and to determine if these are related to *Rlcv* genospecies or perhaps to differences in symbiosis genes. Since the current study is limited to a single semi-controlled greenhouse environment, such markers could potentially be used to assess *Rlcv* community dynamics in the field. Furthermore, group membership could be taken into account in inoculant design, aiming to avoid Generalist strains in favour of Dominants. On the plant genetics side, breeders could select for genotypes that preferentially recruit diverse Dominant populations, which could be provided through inoculation, and they could potentially develop specific faba bean/*Rlcv* Specialist pairs for co-deployment. Such approaches may even allow benefits of inoculation in soils with high background *Rlcv* populations, where it is challenging to improve upon already relatively high nitrogen fixation levels[19,20,32]. Optimisation of faba bean−*Rlcv* interactions can now proceed based on the findings presented here and studies of additional legume-rhizobium relationships can determine if similar principles apply in other systems.

Methods

Soil sampling and analysis

GPS coordinates from each field were recorded, and composite soil samples were taken. All composite soil samples were air-dried prior to analysis. Agricultural soil analysis for phytonutrients standard tests (4505 and 4522) to obtain pH (Rt), mineral nitrogen (Nmin), phosphorus, potassium, magnesium, organic matter, and soil class (JB) were performed by AGROLAB GmbH, Germany. Each soil sample was analyzed to determine the MPN of indigenous rhizobium strains capable of forming nodules on faba bean genotypes Hedin/2 and Melodie, following the methodology described in[22]. Final MPN were determined using tables from ref. 33.

Trapping *Rhizobium leguminosarum* complex sv. *viciae* (*Rlcv*) isolates

14 cm^3 square pots were filled with a pre-sterilized mixture of 3:1 leca:vermiculite and subsequently mixed with 100 g of each soil sample in each pot. Seeds of six faba bean genotypes (Hedin/2, Melodie, Alameda, ILB938/2, VF172/3 cv, and Giza402Gö) were used to trap the *Rlcv* strains (Supplementary Fig. 7a, b). After 8 weeks, nodules were harvested and surface-sterilized with 95% ethanol for 10 s and transferred to 2% sodium hypochlorite for 5 min. Each nodule was allocated to specific positions in a 96-well plate (Supplementary Fig. 7c) and crushed with custom-designed, 3D-printed 12-pestle devices that fit perfectly into each well of the plate, facilitating the simultaneous crushing of multiple nodules (Supplementary Fig. 7d). Using a 12-channel multichannel pipette, 100 µL of sterile water was dispensed into each well and mixed. 40 µL of the resulting mixture were transferred to Rhizobium Defined Media (RDM)[34] to minimize the growth of non-rhizobial isolates. Following the purification of the cultures, a high-throughput DNA extraction was carried out on 96-well plates following the alkaline PEG200 method described in[24]. To confirm that the isolates were *Rlcv* strains, PCRs targeting the *nodD* gene were carried out following the methodology described in ref. 35 with NBA12 and NODDRL2′ primers (Supplementary Data 8) (Supplementary Fig. 7e). Subsequently, ERIC PCRs were carried out following the methodology described in ref. 23 using ERIC1R and ERIC2 primers (Supplementary Data 8) to generate a DNA fingerprint profile for each isolate. A ChemiDoc System was used to visualise and analyse DNA fragments in agarose gels (Supplementary Figs. 7f and 8). *Rlcv* strains with different DNA fingerprint profiles were selected to create the final strain library. We tested the intrinsic resistance of the final rhizobia

libraries, grown on selective media plates, against tetracycline (2 μg mL^{-1}), gentamicin (20 μg mL^{-1}), and neomycin (40 μg mL^{-1}). We found that none of the new isolates were resistant to gentamicin (Supplementary Fig. 7g). The final strain library was stored at −80 °C in glycerol stocks.

Plasmid-ID library

Following the methodology described in ref. 25, a new version of the Golden Gate level-1 backbone vector pOGG026 (RK2-based, broad-host range, lower copy number, and stable in the absence of antibiotic selection) was constructed using the gentamicin-resistance gene pLVC-P2-gent (pOGG009) to obtain the backbone vector PMC-03961 (pL1V-Lv1-gent-RK2-par-ELT4) and create a Plasmid-ID library (Supplementary Data 9) containing 475 plasmids based on the Plasmid-ID system from ref. 24. This system includes a PsnifH promoter driving sfGFP expression for nitrogenase activity in root nodules and unique 12-nucleotide, error-correcting barcodes (ID) to monitor the competitiveness of multiple strains simultaneously. Following the methodology described in ref. 4 competent *E. coli* ST18 cells were used for the plasmid transformations supplemented with 5-aminolevulinic acid (50 μg mL^{-1}) and gentamicin (20 μg mL^{-1}). To validate the performance of the new Plasmid-ID version, it was conjugated into *R. leguminosarum*[36], and using a confocal microscope, we confirmed the activity of the nifH reporter in the nitrogen fixation zone of faba bean nodules (Supplementary Fig. 9).

Confocal microscopy

Confocal microscopy of faba bean nodules was performed using Zeiss LSM780 confocal microscope. The following excitation/emission (nm) settings were used: (i) autofluorescence of cell components 405/410-490, (ii) GFP 488/490-540. The nodules were cut into 100 μm thick sections using Leica VT1000S vibratome.

Tagged *Rlcv* strains with Plasmid-IDs

From the final *Rlcv* strain library, we selected 452 unique strains and conjugated them with the Plasmid-ID library using high-throughput conjugation in 96-well microtiter plates following the methodology described in[24]. Each product of the conjugations was plated onto RDM plates supplemented with gentamicin (20 μg mL^{-1}). Single colonies were re-grown in 96-deep-well plates with RDM liquid media supplemented with gentamicin (20 μg mL^{-1}) in a shaking incubator at 200 rpm and 28 °C for 2 days. To eliminate possible spontaneous resistance to relevant antibiotics, positive plasmid acquisition was verified by colony PCR using GFP_Plasmid_ID-FW and GFP_Plasmid_ID-RV (Supplementary Data 8). The final tagged library comprises 399 strains described in (Supplementary Data 9).

Multi-strain inoculum

To ensure a comparable number of cells from each of the 399 tagged *Rlcv* strains (Supplementary Data 9) and consequently reduce bias in the competition assay, we cultivated the selected strains from our rhizobia library in five different 96-deep well plates. We verified the bacterial cell count of the strain library using optical density (OD). This involved culturing serial dilutions of each strain on plates and converting spectrophotometer readings of culture samples from each strain to cell density (Fig. 1f)[22]. We fed the OD measurements to a script for the OT-2 pipetting robot to transfer the required volume from each well to create the multi-strain inoculum with 2×10^4 cells/ml per strain (Fig. 1g).

Plant assays under greenhouse conditions

Faba bean plants were grown in 14 cm^3 square pots filled with a 3:1 mix of leca:vermiculite. Each pot had individualised irrigation systems to minimize cross-contamination. Nutrient solutions were prepared by combining 398 L of macronutrients with 2 L of micronutrients. For the full fertilization treatment, we used a recipe that included Macronutrients (+N). For treatments without nitrogen (N-free), we used a different recipe that consisted of Macronutrients (-N), described below:

Macronutrients (+N): 25 kg of NPK (14-3-23) +Mg mix was dissolved in 398 L of water. The approximate percentage of each macronutrient was as follows: Nitrate ($N-NO_3$): 10.40%, Ammonium ($N-NH_4$): 3.60%, Phosphorus (P): 2.90%, Potassium (K): 23%, Magnesium (Mg): 3%, Water-soluble sulfur (S): 3.90%, Chloride (Cl): Max 0.05%, and Fluoride (F): Max 0.05%.

Macronutrients (-N): The following nutrients were dissolved in 398 L of water: 4.8 L of sulfuric acid (H_2SO_4) 96%, 11.2 kg of potassium sulfate (SOP), 4.4 kg of magnesium sulfate (16% magnesium oxide, 32% sulfur trioxide), and 3.6 kg of monopotassium phosphate (KH_2PO_4).

Micronutrients: Bought as a liquid mix and chelated with DTPA/EDTA. The approximate percentage of each micronutrient was as follows: Boron (B): 0.23%, Copper (Cu): 0.14%, Iron (Fe): 1.32%, Manganese (Mn): 0.50%, Molybdenum (Mo): 0.05%, and Zinc (Zn): 0.18%.

Each plant was logged into a database and labelled with a barcode to monitor its development and track the harvesting material at the end of the cycle. Plants were harvested after 60 days. Roots of all three treatments were cleaned to check for the presence or absence of nodulation. Roots of inoculated plants were exposed to blue safe light to validate that nodules correspond to labelled strains by observing the detection of GFP (Supplementary Fig. 10). The plant shoots were placed in paper bags placed in drying chambers for at least 3 days.

DNA extraction from nodules

We harvested all root nodules from each individual inoculated plant, surface-sterilised them with 95% ethanol for 10 s, and transferred them to 2% sodium hypochlorite for 5 min. All pooled nodules per plant were placed in 5 ml Eppendorf tubes, and DNA extraction was carried out following the alkaline PEG200 method described in ref. 24. Plant tissue was precipitated at 1000 rpm for 10 min. The supernatant was transferred in aliquots of 30 μL to 96-well PCR plates and stored at −20 °C for future PCR reactions.

Multiplex sequencing

For a two-step PCR, we designed multiplex primers (Supplementary Data 8) by adding the Illumina sequencing primer and flow-cell adapters to sequence our amplicon of interest using Next-Generation Sequencing (NGS). We standardised DNA template concentrations prior to PCR to use the same amount of starting material. All PCRs were run with Q5® High-Fidelity DNA Polymerase from NEB, with limited cycles to minimise the introduction of PCR-generated errors. For the 1st PCR, we followed the NEB Q5® master mix 2× reaction setup and thermocycling conditions provided by NEB (22 cycles with a Tm of 64.5 °C). PCR products were run on a 1.7–2% agarose gel to check the correct band size, and the concentration of the final product was checked with the DNA Qubit fluorescence quantification kit from Thermo Fisher Scientific Inc.

For the 2nd PCR, we used the Nextera XT DNA Library Preparation Kit, which includes i7 and i5 primers. The PCR total reaction volume was 10 μl, which included: 5 μl of NEB Q5® master mix 2×, 1 μl of primer i7, 1 μl of primer i5, 2 μl of 1st PCR product (~1 ng/μl), and 1 μl of DNA-free water. We followed the thermocycling conditions provided by NEB (10 cycles with a Tm of 64.5 °C). 2nd PCR products were run on a 1.7–2% agarose gel to check the correct band size. All samples of each library were pooled in a single tube and cleaned with AMPure XP Bead-Based Reagent following the provided protocol.

We measured the concentration of the final library with the DNA Qubit fluorescence quantification kit from Thermo Fisher Scientific Inc. and diluted it to 10 ng/μl. Each library product was quantified with a Bioanalyzer High Sensitivity DNA Analysis to verify the purity of the products and their final concentration as quality control before the NGS. We denatured and diluted our library following the Illumina

guidelines, and the final multiplex libraries were subjected to NovaSeq 6000 sequencing in PE150 at Novogene Co., Ltd.

Statistical analyses

All statistical analyses were performed in R version 4.2.1[37]. Supplementary Table 2 shows the full list of the R packages used in our analyses. All R scripts used in this study are deposited on GitHub[38].

Rhizobium isolate profiling

Isolate profiling of the nodules was performed using two parameters: relative abundance (RA) and presence/absence. To prevent any bias that can result from the inequal representation of the isolates in the inoculum, we normalised the counts in the nodules (Supplementary Data 5) with the counts in the inocula (Supplementary Data 4) by dividing the former by the latter. In brief, we first performed total sum scaling for both sequencing data sets (so that the library sizes sum to 1). Then we divided these scaled counts in each nodule sample according to their inoculum, multiplied by 1000 and rounded to integers (Supplementary Data 6). To visualise the abundance patterns more clearly, we also performed $\log_2$ transformation with a pseudo-count of 1 (note that in Fig. 2b, c, even though the x-axes show the non-$\log_2$ transformed values, the data points were placed as they were $\log_2$ transformed. This was done with scale_x_continuous function from ggplot2 with the parameter trans = 'log2').

The classification of the isolates into four groups was based on their average RA and occurrence values (Supplementary Fig. 2). First, we simply divide these values with respect to their median values to generate these groups. Next, we kept the median threshold for the occurrence and increased it to 60th percentile for the RA. We did all the analyses for both groupings and generated very similar results (see Supplementary Note). We prefer to present the results based on the one where we used 60th percentile for RA for the following reasons; first, one of the strains showed mostly the characteristics of a Generalists but its RA was a bit higher than that of a Generalists when the median threshold was preferred. Second, when we used median threshold for RA, several Transients moved to Specialist group. This was not very suitable for our assumption that Specialists must be very high abundance, hence we preferred a higher RA threshold (i.e. 60th percentile). Nevertheless, the main conclusions that can be drawn from the study are almost identical for both classifications.

Niche breadth was calculated according to Pandit et al.[26].

$$\text{Breadth}_j = \frac{1}{\sum_{i=1}^{N} P_{ij}^2} \tag{1}$$

where

$$P_{ij} = \frac{\text{RA of isolate } j \text{ in trial } i}{\text{Sum of RAs of isolate } j \text{ in all trials}} \tag{2}$$

This metric evaluates the uniformity of the distribution of the species through the resource states. In our case, the species are the *Rhizobium* isolates, and the resource states are the faba bean plants. Therefore, this metric adds a third dimension to the characterisation of the colonisation groups. Particularly, even though it is expected the Dominants had larger niche breadth in comparison to the remaining groups because this metric can differentiate the species that have the same overall abundance with distinct distributions (such as a uniform or a clumped distribution), we could find differences within the groups with the use of this method.

Interactions between *Rhizobium* isolates

The interaction between the isolate pairs were evaluated by means of their co-occurrence, mutual exclusion, abundance correlation, and their overlap-dissimilarity relationship (i.e. dissimilarity-overlap curve, DOC). Co-occurrence and mutual exclusion calculations were performed on the presence-absence data with the following equations:

$$\text{Cooccurrence}_{ij} = \frac{\text{Number of samples where both isolates } i \text{ and } j \text{ occur}}{\text{Number of samples where only isolate } i \text{ occurs} + \text{Number of samples where only isolate } j \text{ occurs}} \tag{3}$$

$$\text{Mutual exclusion}_{ij} = \frac{\text{Number of samples where only isolate } i \text{ occurs} - \text{Number of samples where only isolate } j \text{ occurs}}{\text{Number of samples where isolates } i \text{ and } j \text{ both occur}} \tag{4}$$

Co-occurrence was based on the number of the samples where a pair of isolates occurred together, divided by the sum of the samples where only one of the isolates occurred. For mutual exclusion, we normalised it (numerator) by co-occurrence (denominator). Further, mutual exclusion has a directionality, i.e. one of the isolates can be competitively superior against the other one, therefore, we took the positive value for this parameter for a pair of isolates (i.e. comparing species i –> species j and species j –> species i *where ME(i,j) = -ME(j,i)*). The significant interactions were then visualised using the igraph package[39]. To filter out the insignificant interactions, we only took the top 10% of the co-occurrence or mutual exclusion values. For co-occurrence, since the values of Specialists were very low, we used the threshold of Generalists (i.e. 90th percentile of the Generalists' co-occurrence distribution, Supplementary Fig. 3a, c, e). For mutual exclusion, the values of Dominants and Generalists were very low, therefore, the universal threshold for this was that of Specialists (i.e. 90th percentile of the Specialists' mutual exclusion distribution, Supplementary Fig. 3b, d, f).

The correlation analyses were based on the $\log_2$-transformed RAs of the isolates with the addition of a pseudo-count of 1, as described above. The Pearson correlation between each isolate pair was calculated using the corr.test function from the psych package. *P*-values were adjusted following the Benjamini–Hochberg method.

DOC analyses[27] were performed using the DOC package (https://github.com/Russel88/DOC). Prior to the analysis, the counts were rarefied using the rarefyFilter function from the seqtime package[40] to 1000 and samples with lower than 1000 counts were discarded. DOC analyses were run with 100 bootstraps. To compare different DOCs, we implemented the measure f_{ns}[27] which is the fraction of data points for which the DOC displays a negative slope. It is formulated as follows:

$$f_{ns} = \frac{\text{number of sample pairs with } O > O_c}{\text{total number of sample pairs}} \tag{5}$$

where O is the overlap and O_c is the changing points where the negative slope begins.

Analysis of plant growth

To determine the relationship between plant biomass and the rhizobial community, we first fitted linear-mixed models (LMMs) with the following equation:

$$\text{Biomass} \sim \text{RA} + (1|\text{PG}) + (1|\text{Batch}) + \text{MDS2} + \text{MDS3} + \text{MDS4}$$

where Biomass is the plant dry weight, RA is the relative abundance of a strain, PG is plant genotype, Batch is the eight batches in the experimental setup, and MDS2-4 are the dimensions from the MDS analysis.

Since the isolate RAs differ largely in maximum values, we first calculated the *z*-scores of the $\log_2$-transformed RAs, so that each isolate had the same average and standard deviation values (0 and 1, respectively). The model included the community structure in terms of MDS2-4 (Supplementary Fig. 4). The MDS analysis was performed on the Cao distances[41] with the vegdist function from vegan package[42] and cmdscale function from base R. Here we did not include the first dimension from the MDS analysis as it was confounded with the batch effect, which was already included in the model as a random effect. The models were fitted using the lmer function from the lmerTest package[43], which produced *P*-values for the fixed effects that were then adjusted for multiple testing with Benjamini–Hochberg method.

We used alpha diversity measures including Shannon's diversity, Simpson's diversity, Fisher's diversity, and evenness[44] for the evaluation of community-biomass relationships. Shannon's diversity, Simpson's diversity, and Fisher's diversity were calculated using estimate_richess function from phyloseq[45]. Evenness was estimated using the sheldon function from the seqtime package[40], according to the following formula:

$$S = \frac{e^H}{N} \tag{6}$$

where H is the Shannon's diversity and N the species number. S ranges from 0 to 1. The distinction between evenness and Shannon's diversity is that the latter considers species count but evenness is independent of it. Hence, even if a small number of isolates are evenly distributed, evenness can be high. After calculating a diversity measure for three distinct colonisation groups, we implemented the following mixed model to find its association with the plant biomass:

$$\text{Biomass} \sim \text{DCG} + (1|\text{PG}) + (1|\text{Batch}) + \text{MDS2} + \text{MDS3} + \text{MDS4}$$

where Biomass is the plant dry weight, D_{CG} is the diversity of a colonisation group, PG is plant genotype, Batch is the eight batches in the experimental setup, and MDS2-4 are the dimensions from the MDS analysis (Supplementary Fig. 4). Again, diversity was normalised (average of 0, standard deviation of 1) prior to the analyses.

Prediction analysis

We performed two types of prediction analyses; one is for the prediction of the plant biomass using the community information as the predictors. The second is the prediction of the community information using plant genetic data. Prediction analyses were performed using the caret package[46] and the ranger method[47], an implementation of the random forest machine learning algorithm[48].

The first analysis was based on the question if including the community information in the random forest model could increase the prediction accuracy of plant growth. Thus, we compared the prediction accuracies from our null model (plant genotype + batch) with a model including the community information (diversity + plant genotype + batch). Diversity information was included in the model as three separate predictors (for the Dominants, the Generalists, and the Specialists). The predictors for the plant genotype were the first ten dimensions of the principal component analysis of the genomic relationship matrix (GRM) based on the faba bean genotype data (Supplementary Data Table 3). For this, we computed GRM following the method proposed by VanRaden[49], implemented through the custom script developed by Moeskjær et al.[29]. This GRM provided a quantitative measure of genetic similarity between the individual plants in our study. Subsequently, we applied PCA to the GRM using the prcomp function in R. We also tested another model with the permuted diversity information.

The prediction part pertaining to the first type of prediction analysis was done as follows: using the createDataPartition function from caret, we performed a hundred random 80%–20% train-test splits in a balanced manner with respect to the batch factor. Using the trainControl function from caret, we performed a random search with sixfold cross validation in 2 repeats, then the best model was evaluated on the test data. The accuracy was based on R^2 values. This process was repeated 100 times (for each random train-test split) producing an average value for accuracy of each group of models.

Genomic prediction analysis was performed according to Moeskjær et al.[29]. In Brief, this analysis is based on the feature selection via GWAS runs followed by the evaluation of the prediction accuracy. Both feature selection (i.e. GWAS) and prediction analysis were performed on the average of the phenotype data of interest (e.g. Shannon's diversity or cumulative relative abundance) across the bioreplicates. The train-test split was done as described above. The training (80%) set was then subjected to GWAS analysis. GWAS analyses were performed using the BLINK method[50] provided within the GAPIT package[51]. First 3 PCAs were included in GWAS models and minor allele frequency threshold was 5%. The 200 SNPs with the lowest *P*-value were then used as the predictors. We fitted a random forest model to predict either Shannon's diversity or cumulative relative abundance of the colonisation groups on the training set, and the best performing model was evaluated on the testing set (same as described above). The importance of each marker was estimated on permutation-basis using the varImp function from caret. This process (from GWAS to accuracy calculation) was repeated 100 times (for each random train-test split) producing an average value for accuracy of each group of models. We also performed the same analysis with 200 randomly chosen SNPs (instead of the top 200) to confirm the validity of our predictive analysis.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequencing reads generated in this study have been deposited in the European Nucleotide Archive under accession code PRJEB81474. All data structures required to reproduce the results of this study can be find as a Supplementary Data file. Source data are provided as a Source Data file. Source data are provided with this paper.

Code availability

The scripts used in this study are available at https://github.com/tyakyol/rhizobialStrains3 (https://doi.org/10.5281/zenodo.13837945)[38].

References

1. Cernay, C., Pelzer, E. & Makowski, D. A global experimental dataset for assessing grain legume production. *Sci. Data* **3**, 160084 (2016).
2. Herridge, D. F., Peoples, M. B. & Boddey, R. M. Global inputs of biological nitrogen fixation in agricultural systems. *Plant Soil* **311**, 1–18 (2008).
3. Jayakodi, M. et al. The giant diploid faba genome unlocks variation in a global protein crop. *Nature* **615**, 652–659 (2023).
4. Skovbjerg, C. K. et al. Genetic analysis of global faba bean diversity, agronomic traits and selection signatures. *Theor. Appl. Genet.* **136**, 114 (2023).
5. Boivin, S. et al. Host-specific competitiveness to form nodules in *Rhizobium leguminosarum* symbiovar *viciae*. *New Phytol.* **226**, 555–568 (2020).
6. Young, J. P. W. et al. Defining the *Rhizobium leguminosarum* species complex. *Genes* **12**, 111 (2021).
7. Cavassim, M. I. A. et al. Symbiosis genes show a unique pattern of introgression and selection within a *Rhizobium leguminosarum* species complex. *Microbial Genomics* **6**, e000351 (2020).

8. Kumar, N. et al. Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol.* **5**, 140133 (2015).
9. Mutch, L. A. & Young, J. P. W. Diversity and specificity of *Rhizobium leguminosarum* biovar *viciae* on wild and cultivated legumes. *Mol. Ecol.* **13**, 2435–2444 (2004).
10. Checcucci, A., DiCenzo, G. C., Bazzicalupo, M. & Mengoni, A. Trade, diplomacy, and warfare: the quest for elite rhizobia inoculant strains. *Front. Microbiol.* **8**, 2207 (2017).
11. Bourion, V. et al. Co-inoculation of a pea core-collection with diverse Rhizobial strains shows competitiveness for nodulation and efficiency of nitrogen fixation are distinct traits in the interaction. *Front. Plant Sci.* **8**, 2249 (2018).
12. Burghardt, L. T. et al. Select and resequence reveals relative fitness of bacteria in symbiotic and free-living environments. *Proc. Natl. Acad. Sci.* **115**, 2425–2430 (2018).
13. Rahman, A. et al. Competitive interference among rhizobia reduces benefits to hosts. *Curr. Biol.* **33**, 2988–3001.e4 (2023).
14. Fields, B., Moffat, E. K., Friman, V.-P. P. & Harrison, E. The impact of intra-specific diversity in the rhizobia-legume symbiosis. *Microbiology* **167**, 1051 (2021).
15. Somasegaran, P. & Bohlool, B. B. Single-strain versus multistrain inoculation: effect of soil mineral N availability on rhizobial strain effectiveness and competition for nodulation on chick-pea, soybean, and dry bean. *Appl. Environ. Microb.* **56**, 3298–3303 (1990).
16. Batstone, R. T., Burghardt, L. T. & Heath, K. D. Phenotypic and genomic signatures of interspecies cooperation and conflict in naturally occurring isolates of a model plant symbiont. *Proc. R. Soc. B Biol. Sci.* **289**, 20220477 (2022).
17. Burghardt, L. T., Epstein, B., Hoge, M., Trujillo, D. I. & Tiffin, P. Host-associated rhizobial fitness: dependence on nitrogen, density, community complexity, and legume genotype. *Appl. Environ. Microb.* **88**, e0052622 (2022).
18. Epstein, B. et al. Combining GWAS and population genomic analyses to characterize coevolution in a legume-rhizobia symbiosis. *Mol. Ecol.* **32**, 3798–3811 (2023).
19. Maluk, M. et al. Fields with no recent legume cultivation have sufficient nitrogen-fixing rhizobia for crops of faba bean (*Vicia faba* L.). *Plant Soil* **472**, 345–368 (2022).
20. Jithesh, T. et al. Recent progress and potential future directions to enhance biological nitrogen fixation in faba bean (*Vicia faba* L.). *Plant-Environ. Interact.* **5**, e10145 (2024).
21. Mendoza-Suárez, M., Andersen, S. U., Poole, P. S. & Sánchez-Cañizares, C. Competition, nodule occupancy, and persistence of inoculant strains: key factors in the rhizobium-legume symbioses. *Front. Plant Sci.* **12**, 690567 (2021).
22. Howieson, J. G. & Dilworth, M. J. *Working with Rhizobia* (Canberra: Australian Centre for International Agricultural Research, 2016).
23. De Bruijn, F. J. Use of repetitive (repetitive extragenic palindromic and enterobacterial repetitive intergeneric consensus) sequences and the polymerase chain reaction to fingerprint the genomes of *Rhizobium meliloti* isolates and other soil bacteria. *Appl. Environ. Microbiol.* **58**, 2180–2187 (1992).
24. Mendoza-Suárez, M. A. et al. Optimizing Rhizobium-legume symbioses by simultaneous measurement of rhizobial competitiveness and N_2 fixation in nodules. *Proc. Natl. Acad. Sci. USA* **117**, 201921225 (2020).
25. Geddes, B. A., Mendoza-Suárez, M. A. & Poole, P. S. A bacterial expression vector archive (BEVA) for flexible modular assembly of golden gate-compatible vectors. *Front. Microbiol.* **9**, 3345 (2019).
26. Pandit, S. N., Kolasa, J. & Cottenie, K. Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* **90**, 2253–2262 (2009).
27. Bashan, A. et al. Universality of human microbial dynamics. *Nature* **534**, 259–262 (2016).
28. Verbruggen, E. et al. Mycorrhizal fungi show regular community compositions in natural ecosystems. *ISME J.* **12**, 380–385 (2018).
29. Moeskjær, S. et al. Major effect loci for plant size before onset of nitrogen fixation allow accurate prediction of yield in white clover. *Theor. Appl. Genet.* **135**, 2021.04.16.440135 (2021).
30. Denison, R. F. & Muller, K. E. An evolutionary perspective on increasing net benefits to crops from symbiotic microbes. *Evol. Appl.* **15**, 1490–1504 (2022).
31. Westhoek, A. et al. Conditional sanctioning in a legume-Rhizobium mutualism. *Proc. Natl Acad. Sci. USA* **118**, e2025760118 (2021).
32. Denton, M. D., Pearce, D. J. & Peoples, M. B. Nitrogen contributions from faba bean (*Vicia faba* L.) reliant on soil rhizobia or inoculation. *Plant Soil* **365**, 363–374 (2013).
33. Woomer, P. L. Most probable number counts. *Methods of Soil. Analysis: Part 2 Microbiological and Biochemical Properties* Vol. 5 (John Wiley & Sons, Ltd, 1994).
34. Ronson, C. W. & Primrose, S. B. Effect of glucose on polyol metabolism by *Rhizobium trifolii*. *J. Bacteriol.* **139**, 1075 (1979).
35. Laguerre, G. et al. Typing of rhizobia by PCR DNA fingerprinting and PCR-restriction fragment length polymorphism analysis of chromosomal and symbiotic gene regions: application to *Rhizobium leguminosarum* and its different biovars. *Appl. Environ. Microbiol.* **62**, 2029–2036 (1996).
36. Young, J. P. W. et al. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **7**, 1–20 (2006).
37. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, 2022).
38. tyakyol. *tyakyol/rhizobialStrains3:* Increased diversity of beneficial rhizobia enhances faba bean growth. (2024).
39. Csárdi, G. et al. igraph: Network Analysis and Visualization in R. R package version 1.5.1 (2024).
40. Faust, K. et al. Signatures of ecological processes in microbial community time series. *Microbiome* **6**, 120 (2018).
41. Cao, Y., Williams, W. P. & Bark, A. W. Similarity measure bias in river benthic Aufwuchs community analysis. *Water Environ. Res.* **69**, 95–106 (1997).
42. Oksanen J. et al. vegan: Community Ecology Package. R package version 2.6-4 (2022).
43. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
44. Sheldon, A. L. Equitability indices: dependence on the species count. *Ecology* **50**, 466–467 (1969).
45. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
46. Kuhn, M. Building predictive models in R Using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
47. Wright, M. N. & Ziegler, A. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
48. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
49. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Diary Sci.* **91**, 4414–4423 (2008).
50. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience* **8**, giy154 (2019).
51. Lipka, A. E. et al. GAPIT: Genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).

Acknowledgements
This work was supported by funding from the European Union's Horizon 2020 Programme for Research & Innovation in the ERA-NET Cofund SusCrop grant no. 771134 to S.U.A. for the ProFaba project, part of the Joint Programming Initiative on Agriculture, Food Security, and Climate Change (FACCE-JPI) and supported by the national Danish funder Grønt Udviklings- og Demonstrationsprogram (GUDP) under Miljø- og Fødevareministeriet, Denmark. In addition, there was funding from the Novo Nordisk Foundation grant no. NNF23OC0081220 to S.U.A. for the N2CROP project and from Innovationsfonden Innoexplorer grant no. 2071-00012B to M.M.S.

Author contributions
M.M.S. and S.U.A. designed the research. M.M.S. and M.N. performed the research. T.Y.A. and M.M.S. analysed the data. M.M.S. and T.Y.A. wrote the first draft. S.U.A. edited the manuscript with input from all authors.

Competing interests
M.M.S. and S.U.A. own shares of SymbioMatch ApS. M.M.S., S.U.A., and M.N. have the European patent application No.: 23216703.1 pending. T.Y.A. declares no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-54940-5.

Correspondence and requests for materials should be addressed to Marcela Mendoza-Suárez or Stig U. Andersen.

Peer review information *Nature Communications* thanks Euan James and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.