Systematic variability enhances the reproducibility of an ecological study

Alexandru Milcu[1,2], Ruben Puga-Freitas[3], Aaron M. Ellison[4,5], Manuel Blouin[3,6], Stefan Scheu[7], Thomas Girin[8], Grégoire T. Freschet[2], Laura Rose[9], Michael Scherer-Lorenzen[9], Sebastien Barot[6], Jean-Christophe Lata[10], Simone Cesarz[11,12], Nico Eisenhauer[11,12], Agnès Gigon[3], Alexandra Weigelt[11,12], Amandine Hansart[13], Anna Greiner[9], Anne Pando[6], Arthur Gessler[14,15], Carlo Grignani[16], Davide Assandri[16], Gerd Gleixner[17], Jean-François Le Galliard[10,13], Katherine Urban-Mead[2], Laura Zavattaro[16], Marina E.H. Müller[14], Markus Lange[18], Martin Lukac[19,20], Michael Bonkowski[17], Neringa Mannerheim[21], Nina Buchmann[21], Olaf Butenschoen[7,22], Paula Rotter[9], Rahme Seyhun[19], Sebastien Devidal[1], Zachary Kayler[14,23] and Jacques Roy[1]

[1]Ecotron (UPS-3248), CNRS, Campus Baillarguet, F-34980, Montferrier-sur-Lez, France.

[2]Centre d'Ecologie Fonctionnelle et Evolutive, CEFE-CNRS, UMR 5175, Université de Montpellier – Université Paul Valéry – EPHE, 1919 route de Mende, F-34293, Montpellier Cedex 5, France.

[3]Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), Université Paris-Est Créteil, 61 avenue du Général De Gaulle, F-94010 Créteil Cedex, France.

[4]Harvard Forest, Harvard University, 324 North Main Street, Petersham, Massachusetts, USA.

[5]University of the Sunshine Coast, Tropical Forests and People Research Centre, Locked Bag 4, Maroochydore DC, Queensland 4558, Australia.

[6]IRD, Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), UPMC, Bâtiment 44-45, deuxième étage, bureau 208, CC 237, 4 place Jussieu, 75252 Paris cedex 05, France.

[7]J.F. Blumenbach Institute for Zoology and Anthropology, Georg August University Göttingen, Berliner Str. 28, 37073 Göttingen, Germany.

[8]Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, RD10, 78026 Versailles Cedex, France.

[9]Faculty of Biology, University of Freiburg, Geobotany, Schaenzlestr. 1, D-79104 Freiburg, Germany.

[10]Institut des Sciences de l'Ecologie et de l'Environnement de Paris (UPMC, UPEC, Paris Diderot, CNRS, IRD, INRA), Sorbonne Universités, CC 237, 4 place Jussieu, 75252 Paris cedex 05, France.

[11]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany.

[12]Institute of Biology, Leipzig University, Johannisallee 21, 04103 Leipzig, Germany.

[13]Ecole normale supérieure, PSL Research University, Département de biologie, CNRS, UMS 3194, Centre de recherche en écologie expérimentale et prédictive (CEREEP-Ecotron IleDeFrance), 78 rue du château, 77140 Saint-Pierre-lès-Nemours, France.

[14]Leibniz Centre for Agricultural Landscape Research (ZALF), Institute of Landscape Biogeochemistry, Eberswalder Str. 84, 15374 Müncheberg, Germany.

[15]Swiss Federal Research Institute WSL, Zürcherstr. 111, 8903 Birmensdorf, Switzerland.

[16]Department of Agricultural, Forest and Food Sciences, University of Turin, largo Braccini, 2, 10095 Grugliasco, Italy.

[17]Department of Terrestrial Ecology, Institute for Zoology, University of Cologne, Zülpicher Str. 47b, 50674 Köln, Germany.

[18]Max Planck Institute for Biogeochemistry, Postfach 100164, 07701 Jena, Germany.

[19]School of Agriculture, Policy and Development, University of Reading, Reading, RG6 6AR, UK.

[20]FLD, Czech University of Life Sciences, 165 00 Prague, Czech Republic.

[21]Institute of Agricultural Sciences, ETH Zurich, Universitätsstrasse 2, 8092 Zürich, Switzerland

[22]Senckenberg Biodiversität und Klima Forschungszentrum BiK-F, Georg-Voigt-Straße 14-16, Frankfurt am Main.

[23]USDA Forest Service, Northern Research Station, Lawrence Livermore National Laboratory, Livermore, California 94550 USA.

Corresponding author: Alexandru Milcu, CNRS, Ecotron - UPS 3248, Campus Baillarguet, 34980, Montferrier-sur-Lez, France, email: alex.milcu@cnrs.fr, phone: +33 (0) 434-359-893.

Many scientific disciplines currently are experiencing a "reproducibility crisis" because numerous scientific findings cannot be repeated consistently[1–4]. A new but controversial hypothesis postulates that stringent levels of environmental and biotic standardization in experimental studies reduces reproducibility by amplifying impacts of lab-specific environmental factors not accounted for in study designs[5–8]. A corollary to this hypothesis is that the deliberate introduction of controlled systematic variability (CSV) in experimental designs can increase reproducibility. We tested this hypothesis using a multi-laboratory microcosm study in which the same ecological experiment was repeated in 14 laboratories. Each laboratory introduced environmental and genotypic CSV within and among treatments in replicated microcosms established in either growth chambers (with stringent control of environmental conditions) or glasshouses (with more variable

environmental conditions). The introduction of genotypic CSV increased reproducibility of results in growth chambers but had no significant effect in glasshouses where reproducibility also was lower. Environmental CSV had little effect on reproducibility. This first deliberate attempt at reproducing an ecological experiment with added CSV reveals that introducing genotypic CSV in experiments carried out under controlled environmental conditions with stringent standardization can increase reproducibility by buffering against unaccounted lab-specific environmental and biotic factors that may otherwise strongly bias experimental outcomes.

Keywords: standardization, microcosms, "reproducibility crisis", experimental methods, genotypic diversity, controlled environment, growth chambers, greenhouses, systematic heterogenization

Reproducibility—the ability to duplicate a study and its findings—is a defining feature of scientific research. In ecology, it is often argued that it is virtually impossible to precisely duplicate any single ecological experiment or observational study because complex ecological interactions between the ever-changing environment and the extraordinary diversity of biological systems exhibiting a wide range of plastic responses at different levels of biological organization together make exact duplication unfeasible[9,10]. Although this may be true for observational and field studies, numerous ecological (and agronomic) studies are carried out with artificially assembled, simplified ecosystems and controlled environmental conditions in experimental microcosms or mesocosms (henceforth, "microcosms")[11–13]. Since biotic and environmental parameters can be tightly controlled in microcosms, results from such studies should be easier to

reproduce. Even though microcosms frequently have been used to address fundamental ecological questions[12,14,15], there has been no quantitative assessment of the reproducibility of any microcosm experiment.

Because it reduces within-treatment variability, experimental standardization— the implementation of strictly defined and controlled properties of organisms and their environment—is widely thought to increase both reproducibility and the sensitivity of statistical tests[7,16]. This paradigm has been challenged recently by several studies on animal behavior that suggest that stringent standardization may, counterintuitively, be responsible for generating non-reproducible results[5–7]; the results may be valid under given conditions (i.e., they are local "truths") but are not generalizable[16,17]. Despite rigorous adherence to experimental protocols, laboratories inherently vary in many conditions that are not measured and are thus unaccounted for, such as experimenter, micro-scale environmental heterogeneity, physico-chemical properties of reagents and lab-ware, pre-experimental conditioning of organisms, and their genetic and epigenetic variation. It even has been suggested[5–7] that attempts to stringently control all sources of biological and environmental variation might inadvertently lead to the amplification of these unmeasured variations among laboratories, thus reducing reproducibility. Some studies have gone even further, hypothesizing that the introduction of controlled systematic variation (CSV) among the replicates of a treatment (e.g., using different genotypes for different experimental replicates or varying pre-experimental conditions) should lead to less variable mean response values between the laboratories that duplicated the experiments[6,7]. In short, reproducibility should increase by shifting the variance from among experiments to within them[7]. If this is true, then introducing CSV will increase researchers' abilities to draw generalizable conclusions about

the directions and effect sizes of experimental treatments, while at the same time reducing the probability of detecting statistically significant treatment effects.

To test the hypothesis that introducing CSV enhances reproducibility in an ecological context, we had 14 European laboratories simultaneously run a simple microcosm experiment using grass (*Brachypodium distachion* L.) monocultures and grass and legume (*Medicago truncatula* Gaertn.) mixtures. This experiment measured the effects of the presence of a nitrogen-fixing legume on ecosystem functioning and productivity in grass-legume mixtures ('net legume effect' hereafter), an approach often used in legume-grass binary cropping systems[18,19] and biodiversity-ecosystem function experiments[20,21]. All laboratories were provided with the same experimental protocol, seed stock from the same batch, and identical containers with which to establish microcosms with grass only and grass-legume mixtures. Alongside a control (CTR) with no CSV and containing a homogenized soil substrate (mixture of soil and sand) and a single genotype of each plant species, we explored the effects of five different types of within- and among-microcosm CSV on experimental reproducibility of the net legume effect (Fig.1): 1) within-microcosm environmental CSV (CSV-WE) achieved by spatially varying soil resource distribution through the introduction of six sand patches into the soil; 2) among-microcosm environmental CSV (CSV-AE), which varied the number of sand patches (none, three or six) among replicate microcosms; 3) within-microcosm biological CSV (CSV-WB) that used three distinct genotypes per species planted in homogenized soil in each microcosm; 4) among-microcosm biological CSV (CSV-AB) that varied the number of genotypes (one, two or three) planted in homogenized soil among replicate microcosms; and 6) both environmental and biotic CSV (CSV-WEB) within microcosms that used six sand patches and three plant genotypes per species in each microcosm. In addition, we tested whether CSV effects depended on the level of

standardization within laboratories by using two common experimental approaches ('SETUP' hereafter): growth chambers with tightly controlled environmental conditions and identical soil (eight laboratories) or glasshouses with more loosely controlled environmental conditions and different soils (six laboratories; Extended Data Table 1). We first tested the response to CSV of twelve variables that are used commonly to describe ecosystem functions of plant-soil microcosms (Extended Data Table 2). We then determined how the different types of CSV affected the mean effect size and its standard deviation (SD) within and among laboratories; lower among-laboratory SD implies that the results were reproduced more closely.

Although each laboratory followed the same experimental protocol, we found remarkably high levels of among-laboratory variation in mean values for the majority of response variables and the net legume effect on those variables (Extended Data Figs 1 and 2). For example, the net legume effect on mean total plant biomass varied from 1.31 to 6.72 g dry weight (DW) per microcosm among growth chambers, suggesting that unmeasured laboratory-specific conditions outweighed effects of experimental standardization. Among glasshouses, differences were even larger: mean plant biomass varied by nearly two orders of magnitude, from 0.14 to 14.57g DW per microcosm (Extended Data Fig. 2).

Among-laboratory SD of net legume effect was significantly affected by CSV, SETUP and their interaction (Table 1, Fig. 2a, b and Extended Data Fig. 3). The main effect of CSV was the lower values of among-laboratory SD in the CSV-AB treatment level relative to CTR ($t_{1,45}$ = 1.97, P = 0.054, Extended Data Fig. 4b), indicating increased reproducibility for CSV-AB. The CSV × SETUP interaction was reflected in the result that among-laboratory SD for CSV-AB was significantly lower only for growth chambers ($t_{1,21}$ = 2.40, P = 0.025, Fig. 2a) and not for glasshouses. Assuming that the grand mean (mean of all laboratories and CSV treatment levels)

is the best available estimate of the "true" legume effect, we also assessed how the CSV treatment affected the deviation from the grand mean (Extended Data Fig. 5). We found that of the five types of CSV, CSV-AB (among-microcosm variance in genotypes) differed least from the grand mean and resulted in the most reproducible results (Fig. 2c).

Within-laboratory SD of the net legume effect was only marginally affected by CSV treatment when the analysis was performed on within-laboratory SD from individual variables (Table 1), but this effect was significant when the analysis was performed on the second principal component (PC2) of a PCA analysis that included all twelve response variables (Extended Data Table 3). No significant CSV × SETUP interaction was found (Fig. 3a). However, we did observe a significant SETUP effect (Table 1 and Extended Data Table 3): within-laboratory SD was lower in growth chambers (Fig. 3b). As we observed a tendency for CSV to increase within-laboratory variation, we also analyzed the impact of the most reproducible CSV treatment—CSV-AB—on the statistical power of detecting the net legume effect within individual laboratories. Adding CSV-AB led to a reduction in statistical power (57% in CTR vs. 45% in CSV-AB) that could be compensated for by doubling the number of microcosms per treatment.

We further explored the relationship between within- and among-laboratory SD to determine whether reproducibility was increased by shifting the variation from among to within laboratories. Although the introduction of CSV generally increased within-laboratory SD of the net legume effect (Extended Data Fig. 6), the treatment level with the highest reproducibility (CSV-AB in growth chambers) only exhibited a non-significant trend of higher within-laboratory SD relative to CTR (Fig. 3c). Moreover, a statistical model of among-laboratory SD as a function of within-laboratory SD, SETUP, and CSV treatment did not reveal a significant three-way

(within-laboratory SD × SETUP × CSV) interaction ($F_{5,120} = 0.49$, $P = 0.784$), although in the growth chamber setup the steepest and flattest slopes were for the CTR and CSV-AB treatment levels, respectively (Fig. 3c). Therefore, although the observed trends are in line with the proposed conjecture that adding CSV enhances reproducibility by increasing within-laboratory variability, our results do not provide unequivocally support for it.

Overall, our findings provide compelling support for the hypothesis that introducing CSV in experimental designs can increase reproducibility of ecological studies[5–7]. We also suggest that the relationship between CSV and reproducibility is purely probabilistic and results from the decreased likelihood that microcosms containing CSV will respond to unaccounted lab-specific environmental factors in the same direction and with the same magnitude. In particular, introducing CSV by using multiple genotypes of study species among replicated microcosms appears to be a good strategy to enhance reproducibility because a mixture of genotypes may limit genotype-specific or assemblage-wide responses to lab-specific environmental variation (see also additional discussion in Supplementary Information). This suggestion is in line with mounting evidence that ecological responses differ significantly among genotypes, and that failure to account for genetic diversity leads to spurious results[22–24]. Interestingly, the effectiveness of CSV-AB in increasing reproducibility was higher in growth chambers, and this result amplifies the importance of introducing CSV in designs with stringent environmental standardization. The lack of a significant effect on reproducibility when introducing CSV-AB in glasshouses could be explained by spatially and temporarily more heterogeneous conditions in the glasshouse environment (e.g. light fleck effects, temperature gradients, etc.) that likely added additional random/unsystematic variation within laboratories. We conclude that to increase reproducibility, ecological experiments should include both rigorous standardization and

controlled systematic variability. Although there are multiple causes for the "reproducibility crisis"[4,25,26], here we show that deliberately including genetic variation may be the simplest solution for increasing the reproducibility of ecological studies performed in controlled environments.

References

1. Massonnet, C. *et al.* Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three Arabidopsis accessions cultivated in ten laboratories. *Plant Physiol.* **152,** 2142–2157 (2010).
2. Begley, C. G. & Ellis, M. L. Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
3. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).
4. Parker, T. H. *et al.* Transparency in ecology and evolution: real problems, real solutions. *Trends Ecol. Evol.* **31,** 711–719 (2016).
5. Richter, S. H., Garner, J. P. & Würbel, H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6,** 257–261 (2009).
6. Richter, S. H., Garner, J. P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7,** 167–8 (2010).
7. Richter, S. H. *et al.* Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One* **6,** e16461 (2011).
8. Wolfinger, R. D. Reanalysis of Richter et al. (2010) on reproducibility. *Nat. Methods* **10,**

373–4 (2013).

9. Cassey, P. & Blackburn, T. Reproducibility and Repeatability in Ecology. *Bioscience* **56,** 958–9 (2006).

10. Ellison, A. M. Repeatability and transparency in ecological research. *Ecology* **91,** 2536–2539 (2010).

11. Lawton, J. H. The Ecotron facility at Silwood Park: the value of' big bottle' experiments. *Ecology* **77,** 665–669 (1996).

12. Benton, T. G., Solan, M., Travis, J. M. & Sait, S. M. Microcosm experiments can inform global ecological problems. *Trends Ecol. Evol.* **22,** 516–521 (2007).

13. Drake, J. M. & Kramer, A. M. Mechanistic analogy: how microcosms explain nature. *Theor. Ecol.* **5,** 433–444 (2012).

14. Fraser, L. H. & Keddy, P. The role of experimental microcosms in ecological research. *Trends Ecol. Evol.* **12,** 478–481 (1997).

15. Srivastava, D. S. *et al.* Are natural microcosms useful model systems for ecology? *Trends Ecol. Evol.* **19,** 379–384 (2004).

16. De Boeck, H. J. *et al.* Global change experiments: challenges and opportunities. *Bioscience* (2015). doi:10.1093/biosci/biv099

17. Moore, R. P. & Robinson, W. D. Artificial bird nests, external validity, and bias in ecological field studies. *Ecology* **85,** 1562–1567 (2004).

18. Sleugh, B., Moore, K. J., George, J. R. & Brummer, E. C. Binary Legume–Grass Mixtures Improve Forage Yield, Quality, and Seasonal Distribution. *Agron. J.* **92,** 24–29 (2000).

19. Nyfeler, D., Huguenin-Elie, O., Suter, M., Frossard, E. & Lüscher, A. Grass-legume mixtures can yield more nitrogen than legume pure stands due to mutual stimulation of

nitrogen uptake from symbiotic and non-symbiotic sources. *Agric. Ecosyst. Environ.* **140,** 155–163 (2011).

20. Temperton, V. M., Mwangi, P. N., Scherer-Lorenzen, M., Schmid, B. & Buchmann, N. Positive interactions between nitrogen-fixing legumes and four different neighbouring species in a biodiversity experiment. *Oecologia* **151,** 190–205 (2007).

21. Suter, M. *et al.* Nitrogen yield advantage from grass-legume mixtures is robust over a wide range of legume proportions and environmental conditions. *Glob. Chang. Biol.* **21,** 2424–2438 (2015).

22. Reusch, T. B., Ehlers, A., Hämmerli, A. & Worm, B. Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2826 (2005).

23. Noguera, D. *et al.* Amplifying the benefits of agroecology by using the right cultivars. *Ecol. Appl.* **23,** 515–522 (2013).

24. Hughes, A. R., Inouye, B. D., Johnson, M. T., Underwood, N. & Vellend, M. Ecological consequences of genetic diversity. *Ecol. Lett.* **11,** 609–623 (2008).

25. Nuzzo, R. How scientists fool themselves – and how they can stop. *Nature* **526,** 182–185 (2015).

26. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (2016).

Acknowledgements

This study benefited from the CNRS human and technical resources allocated to the ECOTRONS Research Infrastructures and the state allocation 'Investissement d'Avenir' ANR-11-INBS-0001 as well as from financial support by the ExpeER (grant no. 262060) consortium

funded under the EU-FP7 research program (FP2007-2013). *Brachypodium* seeds were kindly provided by Richard Sibout (Observatoire du Végétal, Institut Jean-Pierre Bourgin, F-78026 Versailles Cedex France) and *Medicago* seeds were supplied by Jean-Marie Prosperi (INRA Biological Resource Centre, F-34060 Montpellier Cedex 1 France). We further thank Jean Varale, Gesa Hoffmann, Paul Werthenbach, Oliver Ravel, Clement Piel and Damien Landais for assistance throughout the study. For additional acknowledgements see Supplementary Information.

Author contributions

A.M. and J.R. designed the study with input from M.B, S.B and J-C.L. Substantial methodological contributions were provided by M.B., S.S, T.G., L.R. and M.S-L. Conceptual feedback on an early version was provided by G.F., N.E., J.R. and A.M.E. Data were analysed by A.M. with input from A.M.E. A.M. wrote the manuscript with input from all co-authors. All co-authors were involved in carrying out the experiments and/or analyses.

Author Information

The authors declare no conflict of interest. Correspondence and request for materials should be addressed to A.M. (alex.milcu@cnrs.fr).

METHODS

All laboratories tried to the best of their abilities to carry out an identical experimental protocol. Whereas not all laboratories managed to recreate precisely all details of the experimental protocol, we considered this to be a realistic scenario under which ecological experiments using microcosms are performed in glasshouses and growth chambers.

295 **Germination**. The seeds from the three genotypes of *Brachypodium distachyon* (Bd21, Bd21-3
296 and Bd3-1) and *Medicago truncatula* (L000738, L000530 and L000174) were first sterilized by
297 soaking 100 seeds in 100 mL of sodium hypochlorite solution at 2.6% of active chlorine and
298 stirred for 15 min using a magnet. Thereafter, the seeds were rinsed 3 times in 250mL of sterile
299 water for 10-20 seconds under shaking. Sterilized seeds were germinated in trays (10 cm depth)
300 filled with vermiculite. The trays were first kept at 4^{o}C in the dark for three days before they
301 were moved to light conditions (300 µmol m^{-2} s^{-1} PAR) and 20/16^{o}C and 60/70% air RH for day-
302 and night-time, respectively. When the seedlings of both species reached 1 cm in height above
303 the vermiculite they were transplanted into the microcosms.

304 **Preparation of microcosms**

305 All laboratories used identical containers (2-liter volume, 14.8-cm diameter, 17.4-cm height).
306 Sand patches were created using custom-made identical "patch makers" consisting of six rigid
307 PVC tubes of 2.5-cm diam. and 25-cm length, arranged in a circular pattern with an outer
308 diameter of 10cm. A textile mesh was placed at the bottom of the containers to prevent the
309 spilling of soil through drainage holes. Filling of microcosms containing sand patches started
310 with the insertion of the "patch maker" into containers. Thereafter, in growth chamber setups,
311 2000 g dry weight of soil, subtracting the weight of the sand patches, was added into the
312 containers and around the tubes of the "patch maker". In the glasshouse setups with different
313 soils, the dry weight of the soil differed slightly (depending on the soil density) and was first
314 estimated individually in each laboratory as the amount of soil we needed to fill the pots up to 2
315 cm from the top. Finally, the tubes were filled with a mixture of 10% soil and 90% sand. When
316 the microcosms did not contain sand patches, the amount of sand contained in six patches was
317 homogenized with the soil. During the filling of the microcosms, a common substrate for

measuring litter decomposition was inserted at the center of the microcosm at 8-cm depth. For simplicity as well as for its fast decomposition rate, we used a single batch of commercially available tetrahedron-shaped synthetic tea bags (mesh size of 0.25 mm) containing 2 g of green tea (Lipton, Unilever), as proposed by the "tea-bag index" method[27]. Once filled, the microcosms where watered until water could be seen pouring out of the pot. The seedlings were then manually transplanted to predetermined positions (Fig. 1), depending on the genotype and treatment. Each laboratory established two blocks of 36 microcosms each, resulting in a total of 72 microcosms per laboratory, with blocks representing two distinct chambers in growth chamber setups or two distinct growth benches in the same glasshouse**.**

Soils

All laboratories using growth chamber setups used the same soil, whereas the laboratories using glasshouses used different soils (see Extended Data Table 1 for the physicochemical properties of the soils). The soil used in growth chambers was classified as a nutrient-poor cambisol and was collected from the top layer (0–20 cm) of a natural meadow at the Centre de Recherche en Ecologie Expérimentale et Prédictive—CEREEP (Saint-Pierre-Lès-Nemours, France). Soils used in glasshouses originated from different locations. The soil used by laboratory L2 was a fluvisol collected from the top layer (0-40 cm) of a quarry site near Avignon, in the Rhône valley, Southern France. The soil used by laboratory L4 was collected from near the La Cage field experimental system (Versailles, France) and was classified as a luvisol. The soil used in the glasshouse experiments L11 and L12 was collected from the top layer (0-20cm) within the haugh of the river Dreisam in the East of Freiburg, Germany. This soil was classified as an umbric gleysol with high organic carbon content. The soil from laboratory L14 was classified as a eutric fluvisol and was collected on the field site of the Jena Experiment, Germany. Prior to the

established of microcosms, all soils were air dried at room temperature for several weeks and sieved with a 2-mm mesh sieve. A common inoculum was provided to all laboratories to assure that rhizobia specific to *M. truncatula* were present in all soils.

Abiotic environmental conditions

The set points for environmental conditions were 16-hour light (at 300 μmol m^{-2} s^{-1} PAR) and 8-hour dark, 20/16 ^{o}C, 60/70% air RH for day- and night-time, respectively. Different soils (for glasshouses) and treatments with sand patches likely affected water drainage and evapotranspiration. The watering protocol was thus based on drying weight relative to weight at full water holding capacity (WHC). The WHC was estimated based on the weight difference between the dry weight of the containers and the wet weight of the containers 24 h after abundant watering (until water was flowing out of the drainage holes in the bottom of each container). Soil moisture was maintained between 60 and 80% of WHC (i.e. the containers were watered when the soil water dropped below 60% of WHC and water added to reach 80% of WHC) during the first 3 weeks after seedling transplantation and between 50 and 70% of WHC for the rest of the experiment. Microcosms were watered twice a week with estimated WHC values from two microcosms per treatment. To ensure that that the patch/heterogeneity treatments did not become a water availability treatment, all containers were weighed and brought to 70 or 80% of WHC every two weeks. This operation was synchronized with within-block randomization. All 14 experiments were performed between October 2014 and March 2015.

Sampling and analytical procedures

After 80 days, the experiments were stopped and all plants were harvested. Plant shoots were cut at the soil surface level, separated into species and dried at 60^{o}C for three days. Roots and the

364 remaining litter in the tea bags were washed out of the soil using a 1-mm mesh sieve and dried at
365 60ºC for three days. Microcosm evapotranspiration rate was measured before the harvesting as
366 the difference in weight changes from 70% of WHC after 48h. Shoot %C, %N, $\delta^{13}C$, and $\delta^{15}N$
367 were measured on pooled shoot biomass (including seeds) of *B. distachyon* and analyzed at the
368 Göttingen Centre for Isotope Research and Analysis using a coupled system consisting of an
369 elemental analyzer (NA 1500, Carlo Erba, Milan, Italy) and a gas isotope mass spectrometer
370 (MAT 251, Finnigan, Thermo Electron Corporation, Waltham, Massachusetts, USA).

371 **Data analysis and statistics**

372 We focused our analyses on the net legume effect—the difference between the equivalent
373 microcosms with and without legumes—as we considered that comparing within- and among-
374 laboratory variation in the effect size of an experimental treatment (here the presence of a
375 legume) was a more realistic test of reproducibility than comparing absolute values of response
376 variables. All analyses were performed using R version 3.2.4[28]. To assess reproducibility we
377 investigated how CSV treatments affected the standard deviation (SD) of the measured variables,
378 with lower among-laboratory SD indicating increased reproducibility. We opted for SD instead
379 of the coefficient of variation because the net legume effect contained both positive and negative
380 values. As a complementary approach to assess the impact of CSV on reproducibility we
381 explored the extent to which the net legume effect was different from the grand mean (pooled
382 across all laboratories, CSV treatments, and two SETUPs) and used a Kruskal-Wallis test on the
383 ranked differences (of all response variables) from the grand mean.

384 Among-laboratory SD was computed from laboratory means for each response variable,
385 CSV treatments and SETUPs (n = 144; 6 CSV levels × 2 SETUP levels × 12 response variables).
386 Some of the twelve response variables are intrinsically correlated, but most did not have

correlation coefficients > 0.5 (Extended data Fig. 7) and were therefore treated as independent variables. To analyze and visualize the relationship between the SDs calculated from variables with different units, all values of SD were centered and scaled [*z*-scored SD = ($SD_{observed}$ − SD_{mean})/SD_{mean}]. The effects of CSV, SETUP and their interaction on among-laboratory SD was tested with a mixed effects model using the "nlme" package[29] as suggested by Zuur et al. (2009)[30]. The statistical model with the lowest AIC for between-laboratory SD included the response variable as a random factor as well as a "varIdent" weighting function to correct for heteroscedasticity resulting for the variable-specific spread of the residuals (R syntax: "model= lme (between-laboratory SD ~ CSV*SETUP, random=~1|varaible, weights=varIdent (form = ~1|variable))". *A priori* planed contrasts between the CTR and the treatment levels with CSV were performed using Welch's *t*-tests on the *z*-scored normalized SDs.

Within-laboratory SDs were analyzed with two approaches. First, to allow for a direct analytic and graphical comparison with the results for among-laboratory SD, we aggregated the within-laboratory SDs by CSV and SETUP for each response variable (n = 144). A model similar to the one we had used for among-laboratory SD was used to assess the impact of CSV and SETUP. With a second approach we analyzed each response variable separately using mixed effect models with "laboratory" as a random factor and a "varIdent" weighting function to correct for heteroscedasticity resulting for the lab specific spread of the residuals (R syntax: "model= lme (net legume effect ~ CSV*SETUP, random=~1|laboratory, weights=varIdent (form = ~1|laboratory))"; n = 84; 14 laboratories × 6 CSV treatments). As within-laboratory SD data allowed us to account for the inherent collinearity of some of the response variables, we further tested the impact of the CSV on the first and second principal components (PC1 and PC2) derived from a principal component analysis ("prcomp" function in R) using scaled and centered

values from all twelve response variables. PC1 and PC2 were then analyzed with the same mixed-effects model used to analyze the within-laboratory SDs from individual variables. The relationship between within- and among-laboratory SD also was analyzed with a mixed-effects model to test for effects on among-laboratory SD of CSV, within-laboratory SD, and their interactions, and with the source response variable entering the model as a random factor.

As we observed a slight increase in reproducibility of within-laboratory SD for the CSV-AB treatment, we further tested the effect of this treatment on the statistical power for detecting the net legume effect in each individual laboratory. This analysis was performed with the "power.anova.test" function as available in the "base" package. We computed the statistical power of detecting a significant net legume effect (using a one-way ANOVA for the legume treatment) for CTR and CSV-AB for each laboratory and response variable. This allowed us to calculate the average statistical power for the two treatments and the extent to which the number of replicate microcosms needs to be increased for CSV-AB to achieve the same statistical power as for CTR.

Additional references for methods

27. Keuskamp, J. a., Dingemans, B. J. J., Lehtinen, T., Sarneel, J. M. & Hefting, M. M. Tea Bag Index: a novel approach to collect uniform decomposition data across ecosystems. Methods Ecol. Evol. 4, 1070–1075 (2013).

28. Team, R. C. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, 2016).

29. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. NLME: Linear and nonlinear mixed effects models. R Packag. version 3.1-122, http//CRAN.R-project.org/package=nlme 1–

433 336 (2016).

434 30. Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. a & Smith, G. M. Mixed Effects Models

435 and Extension in Ecology with R. (2009). doi:10.1007/978-0-387-87458-

436 **Table 1 | Impact of experimental treatments on among- and within-laboratory SD.** Mixed
437 effects table output showing the impact of controlled systematic variation (CSV), experimental
438 SETUP (glasshouse vs. growth chamber) and their interaction on among- and within-laboratory
439 SD. Complementary analyses assessing the impact of experimental treatments on within-
440 laboratory SD can be found in Extended Data Table 3.
441

Source		Among-laboratory SD		Within-laboratory SD	
	df	F	*P*	F	*P*
CSV	5/121	4.58	0.001	2.27	0.052
SETUP	1/121	1195.70	<0.001	9.38	0.003
CSV×SETUP	5/121	2.48	0.035	0.27	0.926

442
443

Figures

Fig. 1 | Experimental design of one block. Grass monocultures of *Brachypodium distachyon* (green shades) and grass-legume mixtures with the legume *Medicago trunculata* (orange-brown shades) were established in 14 laboratories, with the shades of green and orange-brown representing distinct genotypes. Plants were established in a substrate with equal proportions of sand (black spots) and soil (white), with the sand being either mixed with the soil or concentrated in sand columns to induce environmental CSV. Combinations of three distinct genotypes were used to establish biotic CSV. Alongside a control (CTR) with no controlled systematic variability (CSV) and containing one genotype in a homogenized substrate (soil-sand mixture), five different types of environmental or biotic CSV were used as treatments: 1) within-microcosm environmental CSV (CSV-WE) achieved by spatially varying soil resource distribution through the introduction of six sand patches into the soil; 2) among-microcosm environmental CSV (CSV-AE), which varied the number of sand patches (none, three or six) among replicate microcosms; 3) within-microcosm biological CSV (CSV-WB) that used three distinct genotypes per species planted in homogenized soil in each microcosm; 4) among-microcosm biological CSV (CSV-AB) that varied the number of genotypes (one, two or three) planted in homogenized soil among replicate microcosms; and 6) both environmental and biotic CSV (CSV-WEB) within microcosms that used six sand patches and three plant genotypes per species in each microcosm. The "× 3" indicates that the same genotypic and sand composition was repeated in three microcosms per block. The spatial arrangement of the microcosms in each block was re-randomized every two weeks. The blocks represent two distinct chambers in growth chamber setups, whereas in glasshouse setups the blocks represent two distinct growth benches in the same glasshouse**.**

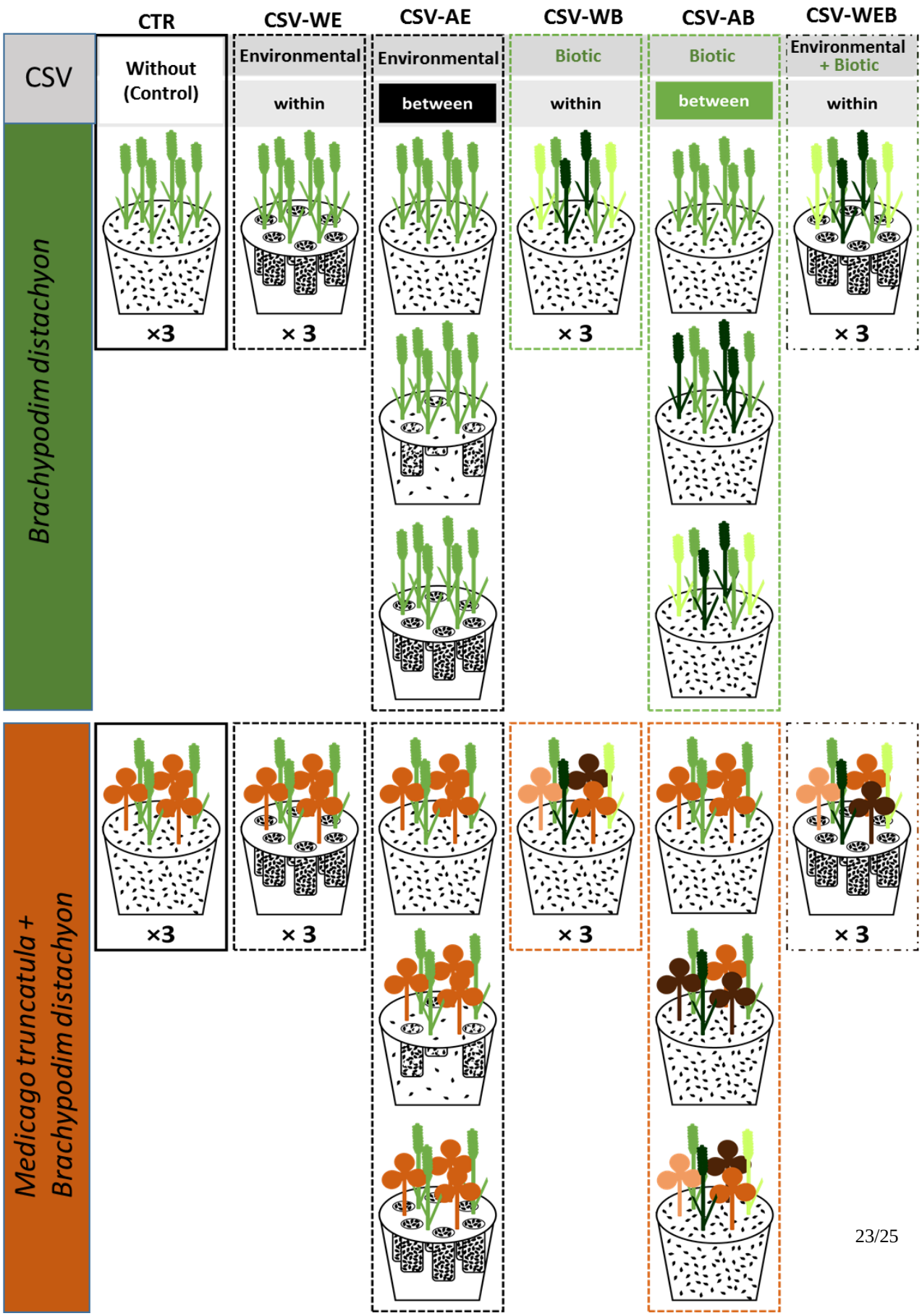

Fig. 2 | Reproducibility as affected by experimental treatments. a, Effects of CSV and SETUP on among-laboratory SD. Lower values indicate enhanced reproducibility. Filled bars and dashed bars represent glasshouse (n = 6) and growth chamber setups (n = 8), respectively. The *P* value shown in **a** represents the result of an *a priori* planned contrast (Welch's *t*-test) between CTR and CSV-AB treatment levels in the growth chamber setup **b,** Overall effect of SETUP on among-laboratory SD. **c**, CSV effects on the mean rank difference from the grand mean (resulted from pooling all 14 laboratories; see Extended Data Fig. 4). Lower difference from the grand mean indicates increased reproducibility. A Kruskal-Wallis test on the ranked differences from the grand mean for the twelve response variables found a significant effect of CSV on the mean ranks ($\chi^2_{5,66}$ = 18.03, *P* = 0.003), with the lowest mean rank difference (*i.e.*, the closer to the grand mean) for the CSV-AB treatment and the highest mean rank difference for the CTR treatment. Bars with error bars represent means ± 1 s.e.m., n = 12 (representing the *z*-scored SD of the twelve measured response variables).

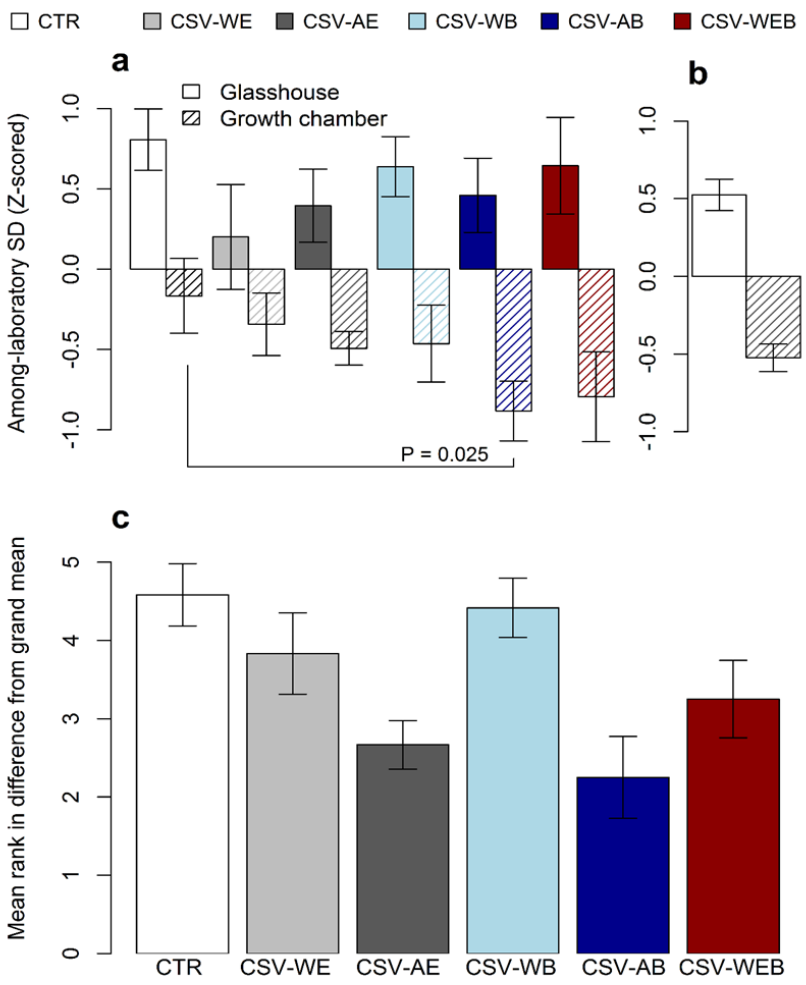

Fig. 3 | Within-laboratory SD of the net legume effect. a, Effects of CSV and SETUP on within-laboratory SD. Filled bars and dashed bars represent glasshouse (n = 6) and growth chamber setups (n = 8), respectively. **b**, Overall effect of SETUP on within-laboratory SD. **c,** Relationship between within- and among-laboratory SD in growth chambers for CTR and CSV-AB. **d,** Relationship between within- and among-laboratory SD in glasshouses for CTR and CSV-AB. The twelve data points per CSV treatment (**c, d**) represent *z*-scored SD of the twelve measured response variables. None of the regressions are significative at $P < 0.05$. Bars with error bars represent means ± 1 s.e.m., n = 12. See Extended Data Fig. 6 for a graph presenting the regression lines for all CSV treatments.