Submitted to: PNAS
Major category: Biological Sciences, **Minor categories:** Ecology, Evolution
Short title: Ecology and evolution of the Linux universe

From computer operating systems to biodiversity: co-emergence of ecological and evolutionary patterns

Authors: Petr Keil[1], Joanne M. Bennett[1,2], Bérenger Bourgeois[3,4], Gabriel E. García-Peña[4,5], A. Andrew M. MacDonald[4,6], Carsten Meyer[1,7], Kelly S. Ramirez[8] & Benjamin Yguel[9,10]

1. German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Deutschland
2. Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany
3. Agroécologie - Agrosup Dijon, Institut national de la recherche agronomique (INRA), Université Bourgogne Franche-Comté - 21000 Dijon - France
4. Centre de Synthèse et d'Analyse sur la Biodiversité - CESAB. Bâtiment Henri Poincaré, Domaine du Petit Arbois. 13857 Aix-en-Provence Cedex 3 – France
5. Facultad de Medicina Veterinaria y Zootecnia. UNAM. México 04510.
6. University of British Columbia, Vancouver, BC, Canada
7. Department of Ecology & Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06511, USA
8. Terrestrial Ecology, Netherlands Institute of Ecology, Droevendaalsesteeg 10, 6708 PB Wageningen, NL
9. Unité MECADEV mécanismes adaptatifs et évolution, UMR 7179 CNRS/MNHN, 4 avenue du Petit Château, 91800 Brunoy, France
10. Centre d'Ecologie et des Sciences de la Conservation (CESCO-UMR 7204), Sorbonne Universités-MNHN-CNRS-UPMC, CP51, 55-61 rue Buffon, 75005, Paris, France

Petr Keil: petr.keil@idiv.de (corresponding author)

Keywords: *Macroecology, cultural evolution, model systems, phylogeny, diversification, power law, macroevolution, niche breadth, log-normal, extinction*

Abstract

Comparisons between biodiversity and other complex systems can facilitate cross-disciplinary exchange of theories and the identification of key system processes and constraints. For example, due to qualitative structural and functional analogies to biological systems, coupled with good data accessibility, computer operating systems offer opportunities for comparison with biodiversity. However, it remains largely untested if the two systems also share quantitative patterns. Here, we employ analogies between GNU/Linux operating systems (distros) and biological species, and look for a number of well-established ecological and evolutionary patterns in the Linux universe. We demonstrate that patterns of the Linux universe match the macroecological patterns: Linux distro commonness and rarity (popularity of a distro) follow a lognormal distribution, power law mean-variance scaling of temporal fluctuation, and there is a significant relationship between niche breadth (number of software packages) and commonness. The diversity in the Linux universe also follows general macroevolutionary patterns: The number of phylogenetic lineages increases linearly through time, with clear per-species diversification and extinction slowdowns, something that is unobservable in biology. Moreover, the composition of functional traits (software packages) exhibits significant phylogenetic signal. Our study provides foundations for using Linux as a model system for eco-evolutionary studies, as well as insights into patterns and dynamics of computer operating systems, which may be used to inform their future development and maintenance. The co-emergence of patterns across systems suggests that some patterns might be produced by system-level properties, independently of system identity, which offers an empirical argument for non-biological explanations of fundamental biodiversity patterns.

Significance statement

Computer operating systems share qualitative properties with biological species -- both undergo evolution, compete for resources, can be classified as common or rare, and have functional traits. Here we found that, apart from the qualitative properties, the two systems also share quantitative patterns: the commonness and rarity are distributed similarly, they fluctuate similarly, rates of origination and extinction slow down over time in a similar way, and in both systems success is linked to the ability to exploit multiple resources. This

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

indicates that quantitative patterns observed in biology can have non-biological explanations. It further suggests that operating systems can serve as a model system for biology, and that patterns found in nature may provide guidance for development of successful operating systems.

Introduction

Exploring analogies between biological and non-biological systems can be beneficial for both natural and social sciences. The comparison of structural analogies across different systems can leverage knowledge on one system to inspire hypotheses and theory for another system. Examples are the adoption of theory first developed in biology to explain the evolution of cultural phenomena (1), and the feedback between theoretical foundations of evolutionary biology, ecology and economics (2–4). Comparing biodiversity with structurally similar systems can help decipher its inherent complexity, by distinguishing phenomena that require biological explanations from those that are produced by system-level constraints, independently of system identity (5). Finally, well-documented non-biological systems might serve as model systems to help address complex questions in biology, or to explore patterns that are effectively unobservable in nature. The latter aspect is potentially crucial for global biodiversity science, given our limited capacity to document biological processes across space, time, and taxa (6, 7), and by our inability to conduct experiments at large spatial and temporal scales.

An area sharing analogical properties with natural systems, and hence offering an opportunity for comparison of patterns, is the universe of open source GNU/Linux operating systems (hereafter Linux, where each operating system is known as a distro; Table 1). Each distro can be viewed as a *species* (8), and the number of devices on which the distro is installed as its *population abundance*. Because distro abundance changes through time, distros have *population dynamics*. Distros contain software packages that determine their applicability in their environments, similar to species *traits*. There is *natural selection* acting in the Linux universe, in that distros compete for resources (adoption by users and attention of developers) (8) and are selected based on their traits. Finally, the evolution of distros is influenced by environmental factors (hardware architecture and user requirements) (9) and constrained by inherited ‘genetic’ material (user accustomization and cost-effective

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

development through reuse of code) (10, 11); as a result, Linux distros have a genealogy which is potentially analogous to a *phylogenetic tree*.

While it has previously been suggested that computing systems might behave analogously to biological systems (8, 12), such analogies have mostly been used to understand the development and maintenance of computing systems (10, 13, 14). However, such comparisons have been mostly *qualitative*, almost anecdotal; thus far, there has been no systematic comparison of *quantitative* laws and patterns that emerge in the two systems.

Despite the general paucity of strict laws in ecology and evolution (15), there are quantitative patterns that are consistently observed across taxonomic, geographical and temporal scales. Familiar examples include: skewed species-abundance distributions (16–18), power-law mean-variance scaling of population fluctuations (19, 20), latitudinal gradients of diversity (21, 22), species-area relationships (21, 23), relationships between niche breadth and range size (24), trait similarities in phylogenetically related species (25, 26) and slowdowns of diversification rates over evolutionary time (27, 28). Typically, these patterns are interpreted in the context of eco-evolutionary processes, although non-biological (statistical, geometrical, neutral) interpretations have also been suggested (29–31).

In this paper, we test if patterns that appear in ecology and evolution also appear in the Linux universe. We look for ecological patterns, such as the skewed species-abundance (or range size) distributions and power-law mean-variance scaling of population fluctuations. We also look for evolutionary patterns, specifically: a phylogenetic signal in traits, slowdowns of diversification rates over evolutionary time, and a positive relationship between niche breadth and range size. Where possible, we discuss the structural properties that likely generate common patterns in both biological and computing systems. Finally, we explore the potential of the Linux universe to serve as a model system for biodiversity by looking at key biological processes and at relationships that are almost impossible to study in nature, such as extinction-speciation dynamics or the speciation-area relationship.

Results and Discussion

We compared the quantitative structure of the two systems and found that the diversity patterns observed in the GNU/Linux universe matched a number of the macroecological and macroevolutionary patterns.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

Macroecological patterns. Linux distro hits per day, HPD (analogous to species commonness) followed a skewed frequency distribution and was well approximated by a log-normal probability density function (Fig. 1A, B), with many distros having low HPD and few distros having high HPD. In ecology, species-abundance distributions (17, 32, 33) and distributions of range sizes (16, 34) are often best described by a skewed log-normal model, which emerges when individuals have a given probability of reproducing or dying, which in turn leads to stochastic multiplicative population dynamics (35). In the Linux universe, similar dynamics likely emerge when users install (reproduction) and uninstall distros (death) on computers.

We also measured the relationship between mean (M) and variance (V) of HPD in 30 distros that were monitored on Distrowatch (www.distrowatch.com) over the last 10 years. We found that the relationship can be modelled by a power function, where log V = -3.34 (±1.25SE) + 2.18 (±0.2SE) log M, and $R^2 = 0.801$ (Fig. 1C). In population biology, such mean-variance scaling is known as Taylor's Power Law (TPL). TPL generally has an $R^2 > 0.8$ in both single- and multi-species systems (19, 29, 36), with the exponent typically falling between 1-2 (20, 37), while values ~2 are common in insect populations (19, 36). Multiple hypotheses have been suggested to interpret empirical TPL exponents (20). Exponents closer to 1 were linked to reproductive asynchrony (38), species interactions (39) or hard upper limit on population size (37). Exponents closer to 2 likely emerge from the simplest multiplicative models such as random walk or deterministic chaos (40). Therefore, the exponent of 2.18 observed in the Linux universe suggests that the temporal dynamics of Linux distro HPD follow a simple unbounded stochastic model, with parameters similar to those found in insect populations.

We detected weak positive, but significant, relationships between HPD and the number of software packages ($R^2 = 0.06$, a = 1.14 (±0.21SE), b = 0.26 (±0.07SE), $p < 0.001$, $r_P = 0.24$; Fig. 2A), and between HPD and the number of all applications ($R^2 = 0.09$, a = 1.65 (±0.07SE), b = 0.14 (±0.03SE), $p < 0.001$, $r_P = 0.302$; Fig. 2B). Yet we found no significant relationship between HPD and the number of specialized software packages (Fig. 2C). Here we consider both the number of packages and the number of applications to be analogous to the ecological concept of niche breadth, which is the suite of environments or resources that a species can inhabit or use (41). By exploiting a greater array of resources and maintaining populations in a wider variety of conditions, a species may become more common, leading to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

a positive correlation between niche breadth and geographical range size (42) - a general ecological pattern (24). We note that the positive relationship between HPD and number of applications can also be analogous to another prevalent ecological pattern known as the *abundance-range size relationship* (41, 42). Our estimated standardized correlations (r_P = 0.24 and 0.302, see above) are roughly in line with (24) who reported values of 0.49 for the correlation between range size and environmental tolerance, and 0.28 for diet breadth vs. range size. Therefore, not only do our findings generalize an ecological pattern found in another scientific field, they also have direct implications for the future development of successful operating systems, since broader functionality is linked to popularity (although this link is weak).

Macroevolutionary patterns. We selected three major Linux families (Debian, Red Hat and SLS) to assess a number of macroevolutionary patterns (Fig. 3A). In all three families, the cumulative number of phylogenetic lineages increased linearly through time (Fig. 3B). In the Debian family, there was a peak in instantaneous per-species speciation rates around 2005, followed by a pronounced slowdown and a peak of extinction rates with a subsequent slowdown in 2006 (Fig. 3C). Similar patterns occurred in Red Hat, but three years earlier (Fig. 3C). In biological systems, diversification slowdowns are pervasive (27, 28), and can be explained by competition for limited resources or for a finite niche space. This reasoning can also explain the diversification slowdowns in the Linux universe: we suggest that in the early 2000s Linux users had become comfortable with certain distros that satisfied most of the potential applications and user requirements – and therefore, speciation slowed down.

Because Linux and software packages have diverse functionalities and connections, we consider two distinct analogies with biological systems: Software packages can either be viewed as functional traits or as symbiotic species. Rather than arguing for either, we examined the Linux universe for both patterns.

First, we made Linux software packages analogous to functional traits, and we found that more closely related distros (measured by temporal distance from their nearest common ancestor) were more similar in their composition of traits (measured by beta-diversity of package composition; Fig. 4). This is surprising, as distros are not confined to directly inherit packages from their ancestors. Instead, creators of a distro are free to load them with any

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

existing open-source software, making the process of package inheritance more similar to, e.g., horizontal gene transfer in bacteria (43). Yet the presence of the signal indicates that some degree of functional similarity with the parent distro is still desired, perhaps by the users (they like the familiar), by the developers (they like to build on what has already been built), and by the need to avoid radical changes that cause bugs (12). This result is consistent with a widespread macroevolutionary pattern known as phylogenetic 'signal' or 'autocorrelation' of traits, i.e. the tendency of closely related species to be functionally more similar than distant relatives (26, 44, 45).

Second, considering the analogy of packages as symbiont species of hosts (distros), we examined the rate at which packages (symbiont) shift between distros (see (46) for such analysis in pathogens of primates): In 10 % (926 out of 9222) of packages, the rate of switches was lower than a neutral pace of evolution (i.e. Brownian Motion) (47), indicating strong host specificity (known as phylogenetic clustering). In the majority of packages however, host shifts occur at a rate that is indistinguishable from Brownian motion. Package niche (host) shifts occurring at slow evolutionary rates is consistent with niche conservatism, a pervasive property of biological systems (26, 45). They can be caused by inherited constraints on potential evolutionary pathways (48) in the form of compatibility constraints on software. Additionally, they can be caused by stabilizing selection against host generalism (45). For example, packaging of software compatible for multiple distros may require time/resources on behalf of developers; and thus a trade-off exists between maximizing compatibility within a distro (fitness) and producing up-to-date, fast evolving software compatible with multiple distros.

'Hard-to-test' macroevolutionary hypotheses. Further, the Linux universe can reveal evolutionary processes for which biological data typically do not exist, since the processes happen at timescales that are impossible to directly observe (e.g., extinction and diversification rates). We have already shown one such process: density-dependent diversification. Another prominent hypothesis is that large areas are more likely to produce new species, since it is more likely for populations to become isolated or to encounter different selection pressures. Isolation is known to limit gene exchange and cause allopatric speciation (21, 49). We found support for the positive area-diversification relationship in the Linux universe – we related the number of diversification events to the human population of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

each country (Fig. 5A), and to the country's area (Fig. 5B), detecting a significant positive relationship in both cases (GLM, quasipoisson family, $p < 0.0001$). If this relationship holds in the Linux universe, a system with numerous structural analogies to biology, then expecting the existence of such relationship in biology (where it cannot be directly observed) seems plausible.

The significance of co-emerging patterns between systems: The Linux universe and biological systems do not only showcase many structural analogies (as already pointed out by (12)), but also share quantitatively similar emerging properties, including patterns of commonness, niche breadth, and evolutionary rates. This has a number of implications. First, it suggests that Linux has the potential to be a model system for macroecology and macroevolution, especially helpful in cases where no biological data exists – we have demonstrated this using direct observation of decreasing speciation and extinction rates over time. However, the potential use of the Linux universe as a model for biodiversity extends beyond the comparisons we make here. More analogies could be made, and patterns analyzed, based on easily retrievable empirical data on Linux. For instance, economic activity tied to Linux applications could serve as an analogy to productivity of geographical areas and allow studying productivity-biodiversity relationships (50). Here we have completed the first step of identifying similarities between the two systems, and these initial analogies should be complemented with a more thorough exploration of the Linux systems to fully assess their utility as a model for biological systems.

At a more basic level the similarities between computing and biological systems have implications for a field commonly called memetics (1), which explains dynamics of cultural phenomena using evolutionary models. Like biological systems, operating systems can be viewed as 'living structures' and as we show, can be viewed as memes. Our key contribution here is that such memes not only obey evolutionary patterns, they also obey ecological patterns, which perhaps calls for an extension of memetics to encompass both cultural evolution and ecology.

Conclusion. Several researchers have already advocated biological evolution and ecological principles as a source of inspiration for software ecology (10) and evolution (9, 10, 12, 51) but, until now, this has remained mostly at the level of qualitative comparisons (the exception being research on networks) (10, 14), and the target audience was mostly computer

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

scientists. Here we show, directly, that simple biological patterns do occur in a universe of computer operating systems, and we argue that these analogies can be useful to ecology, not just to computer science. The unveiled prevalence of simple patterns (e.g., log-normal abundance distributions, phylogenetic signal) in both biology and the Linux universe could mean that these patterns are fundamentally non-biological – a mere outcome of mathematical, geometric, or structural constraints common to both systems (5). Thus, it may be possible that these patterns can be found in other systems where the evolution of its components is constrained by, and based on, the ancestral properties, and is subject to some selection mechanism. However, the logic can also be reversed: There may be more biology in computer operating systems than we would assume from them being just machines – all in all, they are engineered and operated by a biological system, living humans.

Material and Methods

All of the data and R code used for the analyses are available on GitHub at https://github.com/aammd/Linux_diversity_patterns under GNU GP License.

Data on species commonness. We considered popularity (i.e. the number of users) of each distro as a measure of species commonness, which is analogous to either species abundance or range size. However, because GNU/Linux operating systems are mostly freely available for download and use, there is no single way to determine how frequently each distro is used. As a proxy for this information, we used data from Distrowatch (www.distrowatch.com), which provides, among other things, popularity metrics. Popularity on Distrowatch is determined using hits per day (HPD) metric, which is an approximation of public interest in a distro, and we assumed that it is correlated to the actual number of users (i.e. the number of individuals in the population of a Linux distro).

We extracted two datasets: (1) Data on HPD of 275 distros averaged over a yearly period between 27/04/2016 and 28/04/2015 taken from https://goo.gl/?authed=1. These data were used to examine the species-abundance distributions (SAD) (Fig. 1A, B), and to assess the relationship between niche breadth and popularity (Fig. 3). (2) Yearly HPD for each year between (and including) 2002 and 2005 from the main page on www.distrowatch.com (section Page Hit Ranking), downloaded on 28/04/2015. In each year, these data were available only for the 100 most popular distros. This makes them unsuitable for any

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

comprehensive temporal dynamics of SAD, but it allows analyses of population dynamics of the common distros. We used distros that were present in the data for >= 10 consecutive years to assess the temporal mean-variance scaling (Taylor's Power Law; TPL).

Model of species-abundance distribution. We used mean of log(HPD) and standard deviation of log(HPD) as the maximum likelihood estimates of the log-normal probability density function that we plotted in Fig. 1A and B. For panel A, we drew a random sample from that distribution, ordered the outcome, and repeated this 1000 times. We then took the average of the 1000 orderings – this is the red line in Fig. 1A.

Taylor's Power Law (TPL). For each distro with >= 10 years monitored on Distrowatch we calculated log-transformed temporal variance of HPD and log-transformed temporal mean of HPD. This resulted in 30 distro-specific log(Mean) vs. log(Variance) pairs (data points). Although a non-linear regression is potentially better suited to estimate the TPL scaling exponent, we following the practice in the broad literature on TPL and we fitted a normal linear regression of log(Variance) versus log(Mean) using log(Mean) vs. log(Variance) data (equation of the fitted regression in Fig. 1C). We took the slope of the regression as the estimate of TPL scaling exponent.

Data on niche breadth. We used three proxies of niche breadth of Linux distros, which we also interpreted as a position of the distro on the specialist-generalist continuum:

(i) *Number of packages*. We made the analogy between species functional traits and software packages that come pre-installed with each distro, and we used number of software packages as niche breadth, with specialist distros having fewer packages, and generalist distros having more packages. This number of packages was extracted for each of the 275 distros that were listed at the site (specifically at https://goo.gl/?authed=1) for the period between 27/04/2016 and 28/04/2015. On Distrowatch there is a specialized site giving detailed information on each distro. Example is an Ubuntu-specific website at http://goo.gl/997vtZ. We extracted Full Package List for each distro (the Ubuntu-specific example is at http://goo.gl/0Qhflk) and derived numbers of packages from the number of rows in this table.

(ii) *Number of applications*. We used 32 'application' categories defining the broad purpose of each distro. These categories were: Assistive, Beginners, Chromebooks, Clusters, Data Rescue, Desktop, Disk Management, Education, Firewall, Forensics, Free Software,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

Gaming, High Performance Computing, Live CD, Live DVD, Live Medium, Multimedia, Myth TV, Netbooks, Network Attached Storage, Old Computers, Privacy, Raspberry Pi, Router, Scientific, Security, Server, Source-based, Specialist, Telephony, Thin Client, UNIX. We use number of these applications as a measure of niche breadth. We merged Live CD, Live DVD and Live Medium to a single category called Live CD. On Distrowatch, each distro can be labeled with any combination of these categories. These labels can be obtained from each distro-specific website (example of Ubuntu: http://goo.gl/997vtZ).

(iii) *Number of specialized applications*. We used the same number of applications as above, but we excluded the broad 'Desktop' and 'Live CD' categories, so that the number reflects only the relatively narrow applications.

Relationship between niche breadth and commonness. We fitted generalized linear model (quasipoisson family, log link function) with HPD (the commonness) as a response and the three measures of niche breadth (described above) as predictors; we log-transformed the number of packages prior to the modelling.

Phylogenetic information. The majority of current GNU/Linux distros descended from three original distros: Debian, Red Hat and Soft Landing Linux system (SLS, later known as Slackware). Family trees of these three Linux families were compiled by GNU/Linux Timeline project consisting of A. Lundqvist, D. Rodic, M.A. Mustafa, A. Urosevic and J.A. Sandoval; the coded trees are available at http://futurist.se/gldt/ under GNU Free Documantation Licence. We converted the family trees into Newick phylogenetic trees, and for each distro we extracted the date at which its development began ("speciation"), and the date at which development ceased ("extinctions") – information that is virtually impossible to get from biological trees based on molecular sequences and morphological traits.

Speciation and extinction through time. We separated the process of diversification among GNU/Linux distros into speciation and extinction. For each of the three distro families we plotted: (i) The cumulative number of speciation events, which is the number of distros which had been created up to a given date (even if development had ceased) – analogous to cumulative speciation through time plots created for biological systems, the so-called 'lineage-through-time plots' (52). (ii) The cumulative number of extinction events through time, which is the cumulative number of distros which had ceased development – such plots

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

would be desirable in biological systems, but are usually impossible to create due to a lack of data. (iii) Per-species instantaneous speciation and extinction rates; these are numbers of distros that were created or their development ceased in a given month, divided by total number of distros in that month. The per-species instantaneous speciation rate was of key interest, since it is the quantity that is expected to slow down over time.

Data on composition of software packages (traits). We extracted complete lists of software packages contained by default in the last stable versions of 57 Debian-based distros (Fig. 4); we extracted the data from www.distrowatch.com on 14/05/2016 (see the Methods section on niche breadth for more details).

Phylogenetic signal of traits. Using the data described above, we created a package X distro binary matrix, with 1 for presence of the packages, and 0 for absence. We calculated two distance matrices based on compositional dissimilarity between the 57 Debian-based distros; the compositional dissimilarity was measured by β_j (Jaccard beta) and β_{sim} (Beta sim) (53). Using the Debian phylogeny, we also calculated distance matrix based on phylogenetic distance between the 57 distros. We then used Mantel tests (1000 permutations) to detect significant correlations between distance matrices representing compositional dissimilarity and phylogenetic distance. We also calculated Spearman correlations (Rho) between the matrices.

Host specialization of packages. Using 57 Debian-based distros for which we had data on software packages, we classified the packages to host generalists and host specialists, according to the rate of switches between their hosts. From 49,906 packages in our sample, 81 % (40,684) were hosted in one distro (Debian). We excluded all packages that were unique to a single distro, ending up with 9222 packages. We tested for host specialization in each of these packages by comparing the mean pairwise distance between the hosts with the expected mean pairwise distance assuming a null model. The expected mean pairwise distance was calculated by randomly shifting (1000 times) the packages in distros (community) by drawing host from the pool of hosts occurring in the distance matrix (phylogeny pool) with equal probability, i.e. between related hosts. We classified distros with a Z test, using a statistical threshold of $P < 0.05$ for host specialists. The analysis was performed using ses.mpd function from the R package picante (54).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

Relationship between area and diversification. We extracted information on country of origin of each distro from www.distrowatch.com, where each distro has its dedicated website (e.g., the Ubuntu-specific website at http://goo.gl/xzTk), and the header gives the country of origin. For each country we extracted its human population and area from www.geonames.com using R package geonames. We fitted generalized linear model (quasipoisson family, log link function) with number of diversification events as a response and either country human population (log transformed) or country area (log transformed) as predictors.

Acknowledgements

We are grateful to Luke Harmon and Antonin Machac for valuable comments. Ladislav Bodnar from www.distrowatch.com assisted with access to the raw data. This research was supported by the French Foundation for Research on Biodiversity (FRB) through its synthesis centre, CESAB (http://www.cesab.org/) and was conducted at a workshop (25-29/4/2016, Aix-en-Provence, France) organized by Alison Specht (CESAB), together with Marten Winter and the synthesis division (sDiv) of the German Centre for Integrative Biodiversity Research (iDiv). BY benefited from a postdoctoral fund from the French national research agency LabEx ANR-10-LABX-0003-BCDiv, in the context of the “Investissements d'avenir” n° ANR-11-IDEX-0004-02. KSR was supported by ERC Adv grant 26055290.

References

1. Dawkins R (1976) *The selfish gene* (Oxford University Press, Oxford).
2. Malthus TR (1798) *An essay on the principle of population* (J. Johnson, London).
3. Worster D (1994) *Nature’s economy: a history of ecological ideas* (Cambridge University Press).
4. Bonds MH, Dobson AP, Keenan DC (2012) Disease Ecology, Biodiversity, and the latitudinal gradient in income. *PLOS Biol* 10(12):e1001456.
5. Frank SA (2009) The common patterns of nature. *Journal of Evolutionary Biology* 22(8):1563–1585.
6. Hortal J, et al. (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu Rev Ecol Evol Syst* 46(1):523–549.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

7. Meyer C, Kreft H, Guralnick R, Jetz W (2015) Global priorities for an effective information basis of biodiversity distributions. *Nat Commun* 6:8221.

8. Mens T, Maelick C, Grosjean P (2014) ECOS: Ecological studies of open source software ecosystems, pp 403–406.

9. Yan K-K, Fang G, Bhardwaj N, Alexander RP, Gerstein M (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *PNAS* 107(20):9186–9191.

10. Fortuna MA, Bonachela JA, Levin SA (2011) Evolution of a modular software network. *PNAS* 108(50):19985–19989.

11. Myers CR (2003) Software systems as complex networks: structure, function, and evolvability of software collaboration graphs. *Phys Rev E Stat Nonlin Soft Matter Phys* 68(4 Pt 2):46116.

12. Mens T, Claes M, Grosjean P, Serebrenik A (2014) Studying evolving software ecosystems based on ecological models. *Evolving Software Systems*, eds Mens T, Serebrenik A, Cleve A (Springer Berlin Heidelberg), pp 297–326.

13. Pang TY, Maslov S (2013) Universal distribution of component frequencies in biological and technological systems. *PNAS* 110(15):6235–6239.

14. Valverde S, Sole RV (2015) Punctuated equilibrium in the large-scale evolution of programming languages. *Journal of The Royal Society Interface* 12(107):20150249.

15. Lawton JH (1999) Are there general laws in ecology? *Oikos* 84(2):177–192.

16. Gaston KJ (1996) Species-range-size distributions: patterns, mechanisms and implications. *Trends in Ecology & Evolution* 11(5):197–201.

17. McGill BJ, et al. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters* 10(10):995–1015.

18. Morlon H, et al. (2009) Taking species abundance distributions beyond individuals. *Ecology Letters* 12(6):488–501.

19. Taylor LR, Woiwod IP (1980) Temporal Stability as a Density-Dependent Species Characteristic. *Journal of Animal Ecology* 49:209–224.

20. Kendal WS (2004) Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity* 1(3):193–209.

21. Rosenzweig ML (1995) *Species diversity in space and time* (Cambridge University Press).

22. Gaston KJ (2000) Global patterns in biodiversity. *Nature* 405(6783):220–227.

23. Drakare S, Lennon JJ, Hillebrand H (2006) The imprint of the geographical, evolutionary and ecological context on species–area relationships. *Ecology Letters* 9(2):215–227.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

24. Slatyer RA, Hirst M, Sexton JP (2013) Niche breadth predicts geographical range size: a general ecological pattern. *Ecol Lett* 16(8):1104–1114.

25. Darwin C (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (John Murray, London).

26. Losos JB (2008) Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters* 11(10):995–1003.

27. Rabosky DL, Glor RE (2010) Equilibrium speciation dynamics in a model adaptive radiation of island lizards. *PNAS* 107(51):22178–22183.

28. Moen D, Morlon H (2014) Why does diversification slow down? *Trends in Ecology & Evolution* 29(4):190–197.

29. Hubbell SP (2001) *The unified theory of biodiversity and biogeography* (Princeton Univ. Press).

30. Bell G (2001) Neutral macroecology. *Science* 293(5539):2413–2418.

31. Šizling AL, Kunin WE, Šizlingová E, Reif J, Storch D (2011) Between geometry and biology: the problem of universality of the species-area relationship. *Am Nat* 178(5):602–611.

32. Fischer RA (1943) A theoretical distribution for the apparent abundance of different species. *Journal of Animal Ecology* 12:54–58.

33. Preston FW (1948) The commonness, and rarity, of species. *Ecology* 29:254–284.

34. Gaston KJ (2003) *The Structure and Dynamics of Geographic Ranges* (Oxford Univ. Press).

35. Lande R, Engen S, Saether B-E (2003) *Stochastic population dynamics in ecology and conservation* (Oxford University Press, Oxford).

36. Taylor LR, Woiwod IP (1982) Comparative synoptic dynamics. I. Relationships between inter- and intra-specific spatial and temporal variance/mean population parameters. *Journal of Animal Ecology* 51:879–906.

37. Keil P, Herben T, Rosindell J, Storch D (2010) Predictions of Taylor's power law, density dependence and pink noise from a neutrally modeled time series. *Journal of Theoretical Biology* 265(1):78–86.

38. Ballantyne F, Kerkhoff AJ (2007) The observed range for temporal mean-variance scaling exponents can be explained by reproductive correlation. *Oikos* 116(1):174–180.

39. Kilpatrick AM, Ives AR (2003) Species interactions can explain Taylor's power law for ecological time series. *Nature* 422(6927):65–68.

40. Perry JN (1994) Chaotic dynamics can generate taylor's power law. *Proceedings of the Royal Society of London B: Biological Sciences* 257(1350):221–226.

41. Gaston KJ, Blackburn TM, Lawton JH (1997) Interspecific abundance-range size relationships: an appraisal of mechanisms. *Journal of Animal Ecology* 66:579–601.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

42. Brown JH (1984) On the relationship between abundance and distribution of species. *The American Naturalist* 124(2):255–279.

43. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.

44. Harvey PH, Pagel MD (1991) *The comparative method in evolutionary biology* (Oxford University Press, Oxford).

45. Wiens JJ, et al. (2010) Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters* 13(10):1310–1324.

46. Cooper N, Griffin R, Franz M, Omotayo M, Nunn CL (2012) Phylogenetic host specificity and understanding parasite sharing in primates. *Ecol Lett* 15(12):1370–1377.

47. Felsenstein J (1985) Phylogenies and the comparative method. *The American Naturalist* 125:1–15.

48. Cooper N, Jetz W, Freckleton RP (2010) Phylogenetic comparative approaches for studying niche conservatism. *Journal of Evolutionary Biology* 23(12):2529–2539.

49. Lomolino MV, Riddle BR, Whittaker RJ, Brown JH (2010) *Biogeography* (Sinauer Associates, Sunderland, Massachusetts). 4th Ed.

50. Currie DJ, et al. (2004) Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecology Letters* 7(12):1121–1134.

51. Myers CR (2003) Software systems as complex networks: structure, function, and evolvability of software collaboration graphs. *Phys Rev E Stat Nonlin Soft Matter Phys* 68(4 Pt 2):46116.

52. Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. *Phil Trans R Soc B* 344:305–311.

53. Koleff P, Gaston KJ, Lennon JJ (2003) Measuring beta diversity for presence–absence data. *Journal of Animal Ecology* 72(3):367–382.

54. Kembel SW, et al. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26(11):1463–1464.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

496 **Table 1** Summary of analogies between ecological terms and the 'Linux universe', i.e. the
497 system encompassing all Linux distros, Linux software packages, their developers and users,
498 and relationships between those.
499

Ecological terminology	Linux definition
Species	A Linux distribution, most commonly referred to as 'distro'. A distro is a computer operating system comprised of a collection of software packages.
Phylogeny	Evolution of, and relationships between, distros over time. All Linux distros have evolved from three main distros: Debian, Red Hat and SLS.
Commonness (abundance, range size)	Hits per day (HPD). HPD is a yearly average of the number of times per day any given distro page on DistroWatch.com is accessed. HPD is a proxy for distro popularity.
Diversification event	Date at which the development of a distro began.
Extinction event	Date at which development on a distro ceased.
Functional or life-history traits	Software packages available with each distro. These packages or 'traits' determine the applicability of the distro.
Natural selection	Use and popularity of a distro is based on users and downloads. Unused and unpopular distros go extinct.
Area/Productivity	Population of country where distro was developed. Alternatively this could be the user base of each distro, however those statistics are not available.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016

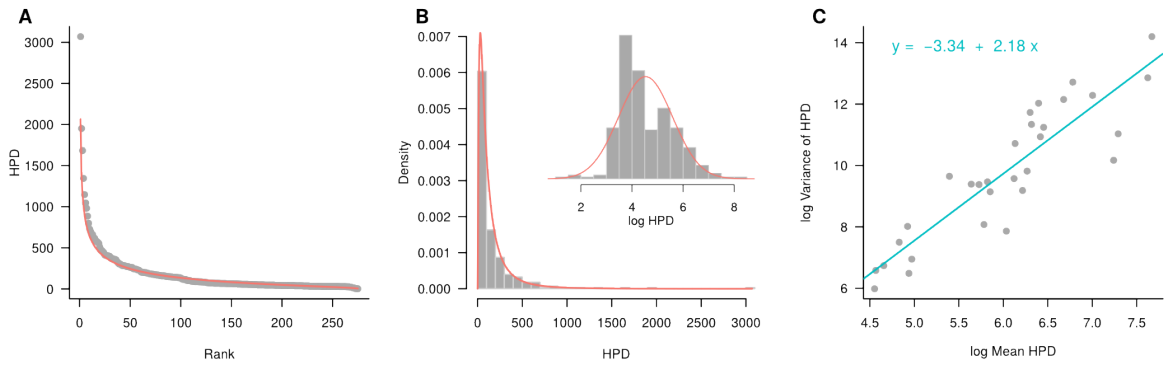


Figure 1 Two macroecological patterns that also appear in the Linux universe. (A-B) Skewed species-abundance distribution (SAD) of Linux distros represented as: (A) species rank abundance curve, (B) species abundance distribution (log-transformed version as inset). Grey points and grey bars are the data, red lines show maximum likelihood fit of a log-normal probability density function. (C) Scaling of temporal variance with temporal mean, also known as Taylor's power law. Blue line (and the equation) is the mean of a linear regression fitted through the log-transformed data.

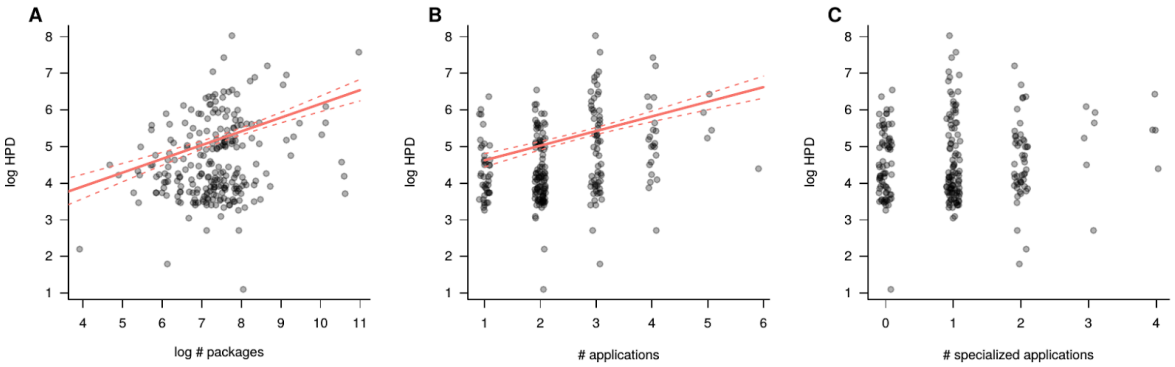
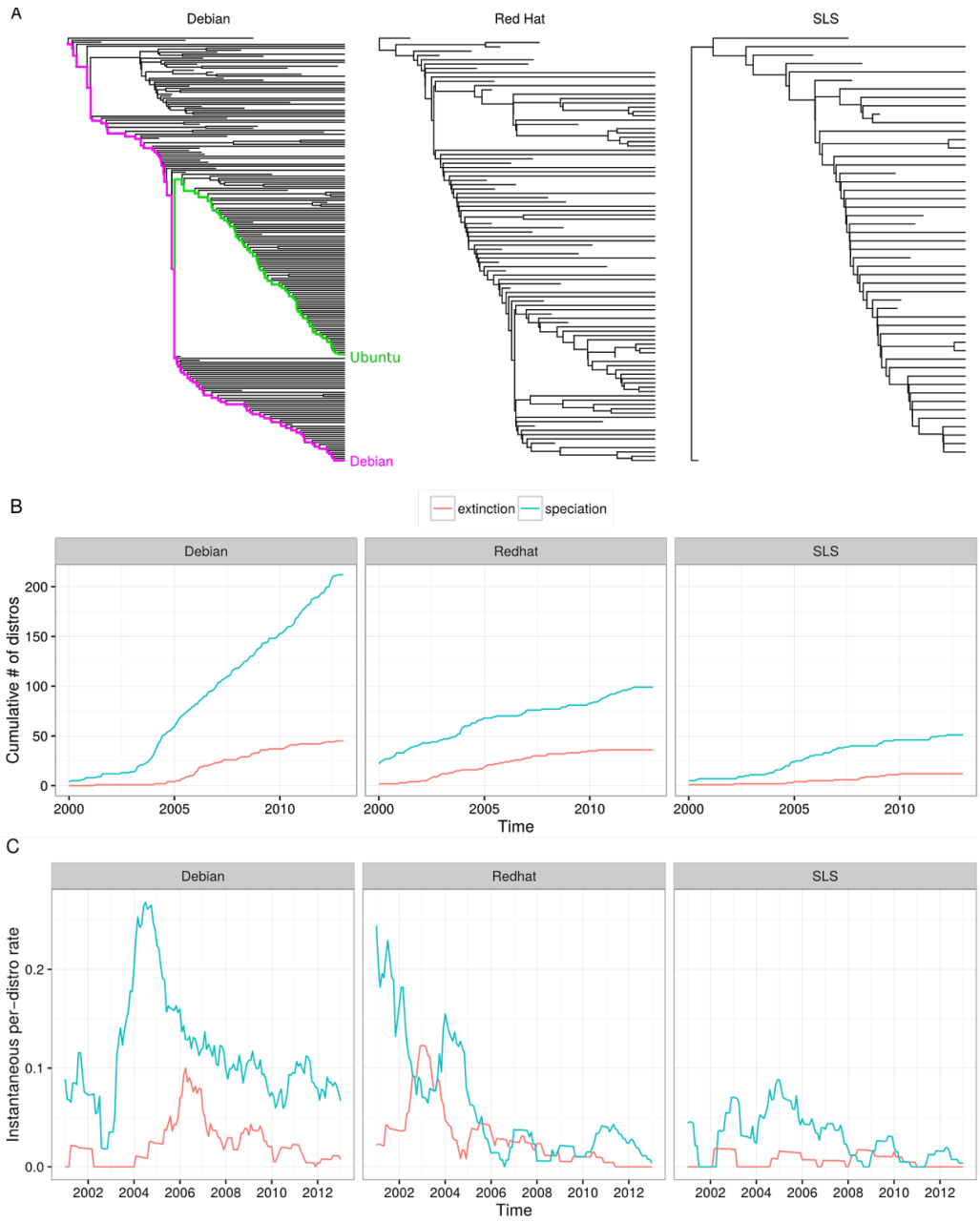


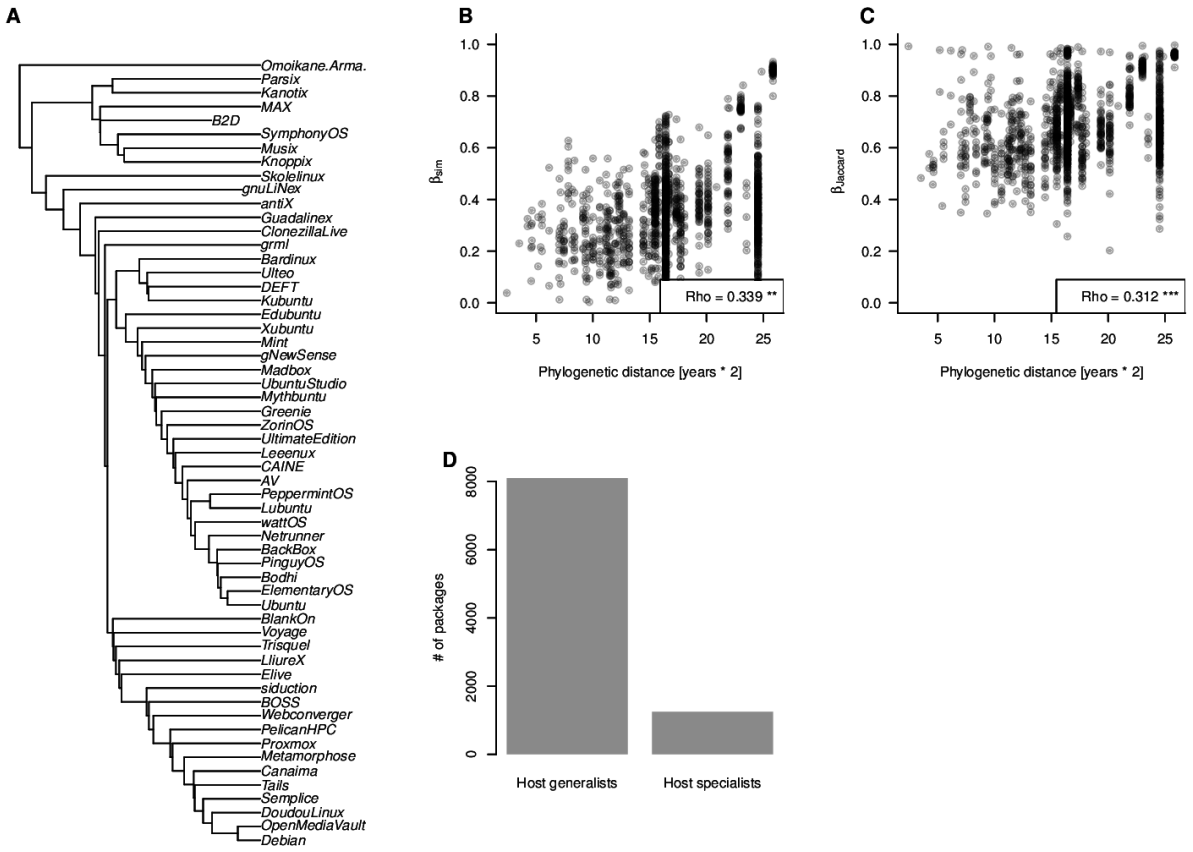
Figure 2 Three measures of niche breadth and their relationship with abundance, measured as HPD. (A) Niche breadth measured as number of packages, (B) niche breadth measured as number of all applications, (C) niche breadth measured as number of specialized applications (not including 'Desktop' and 'Live CD'). Solid red lines are means of quasipoisson log-linear regressions, dashed lines are standard errors of the predicted means.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016



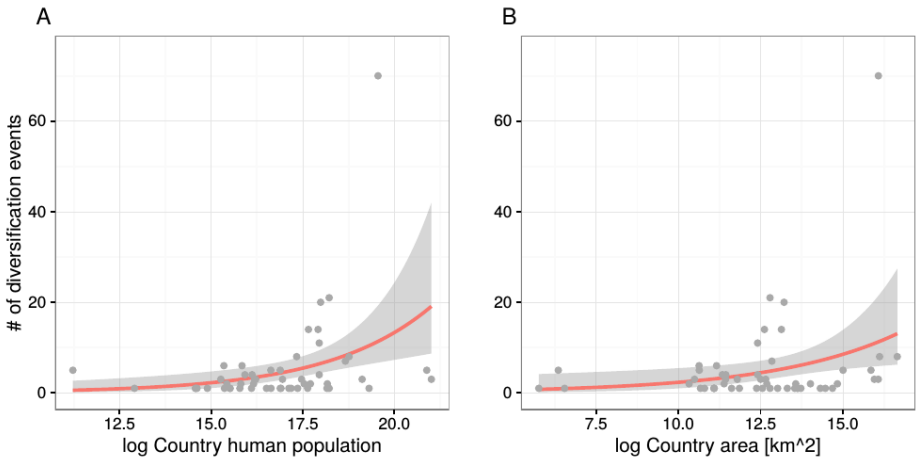
512 **Figure 3** A macroevolutionary pattern in the Linux universe: the diversification and
513 extinction of Linux distros through time. (A) Phylogenetic trees of the three families of Linux
514 distros that were used to make the panels below. Branches of two major distros, Ubuntu and
515 Debian, are highlighted by green and purple. (B**)** Accumulation of new species (distros) and
516 extinctions over time. (C) Instantaneous diversification rate per distro.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016



517 **Figure 4** Phylogenetic signal in the composition of binary traits (sofware packages),
518 expressed as the relationship between pairwise phylogenetic distances and compositional
519 dissimilarity of software packages. (A) Phylogenetic tree of the 57 Debian-based distros used
520 for the analysis. (B-C) Significant positive correlation between phylogenetic distance and
521 trait compositional dissimilarity measured by β_j (B) and β_{sim} (C). Rho is Spearman correlation,
522 asterisks indicate probability that there is no correlation, calculated using Mantel test (9999
523 permutations, Z-score as test statistic), with *** for p<0.001 and ** for p<0.01. (D)
524 Frequency of packages classified as host specialists (they are significantly phylogenetically
525 clustered) and host generalist (not different from what is expected from Brownian motion
526 evolution).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016



527 **Figure 5** Using Linux to investigate 'hard-to-test' macroevolutionary hypotheses: Number of
528 diversification events (i.e. originations of new distros) as a function of (A) population of a
529 country in which distros originate and (B) area of the country. Red lines are mean values from
530 quasipoisson log-linear regression, shaded areas are 95% confidence intervals of the means.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2367v2 | CC BY 4.0 Open Access | rec: 22 Aug 2016, publ: 22 Aug 2016