Editorial

From Galactic archeology to soil metagenomics – surfing on massive data streams

Soil microbiologists make their discovery in the dirt and rarely look at the stars. The words of Leonardo da Vinci 'We know more about the movement of celestial bodies than about the soil underfoot' are as applicable now as they were in 1510 and, currently, scientists deciphering soil microbe genomes and exploring the metagenomes of soil ecosystems may learn from their sky-gazing colleagues. Keeping their feet in the mud, but having their head in the sky, may help them to avoid the (meta)genome-analysis gridlock. Metagenomics involves sampling and sequencing the genome sequences of a community of organisms that inhabit a common environment, such as the ocean, the soil or the human gut (Handelsman, 2004; Hugenholtz & Tyson, 2008). Metagenomics provides an unbiased picture of the community structure (species richness and distribution) and its functional potential. It is rapidly moving from being a description tool to an experimental tool as a result of comparisons now being made of metagenomes submitted to environmental perturbations. Soils are home to microbial communities whose aggregate membership (5×10^{30}) (Whitman *et al.*, 1998) outnumbers the stars shining in the sky (7×10^{21}). Soil complex habitats contain an estimated 10^4 to 10^6 species in a single gram (Gans *et al.*, 2005; Curtis & Sloan, 2006). Once the genomes of these hundreds of thousands of species that crawl on weathered rocks, decaying organic matter and in the rhizosphere are catalogued, the streams of data arising from soil will equal those pouring from star-gazing telescopes. However, a major problem has taken the microbiologist community off-guard: how to analyze this exponentially increasing amount of sequence data. As quoted from Kahvejian *et al.* (2008) 'This surge of new data can be received as a flood, overwhelming the unsuspecting researcher, or as a tremendous wave that can be surfed to new horizons.' Soil (meta)genomics sprang from fast advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. In the coming years, mining a day's worth of data will take more than a day of large supercomputer time, and the fraction of the available data that we will be able to mine and analyze in a useful period of time will rapidly dwindle towards zero. Overcoming this bottleneck requires a shift in our frame of mind from mining databases to mining data streams. Recent papers (Singh *et al.*, 2009) have highlighted the needs, requirements and challenges for sequencing microbial genomes and soil metagenomes in line with the Human Microbiome Project (http://nihroadmap.nih.gov/hmp/) and the Global Ocean Sampling Expedition (Rusch *et al.*, 2007). Before discussing how other scientific communities are navigating through the hazards of massive data streams, we will summarize the current status of genomics and metagenomics of soil micro-organisms.

Unraveling genomes of subterranean species

As of today, microbiologists have unraveled the genomes of more than 1000 bacterial and 100 fungal genomes (http://genomesonline.org/), and more than 4000 metagenomes are stored in the MG-RAST database (Meyer *et al.*, 2008). Next-generation sequencing machines, such as Roche 454 GS FLX and Illumina Solexa GA, can now run through a bacterial and fungal genome in a single day, and are busily churning out multiple sequences from an ever-expanding list of species. Through large-scale sequencing initiatives, such as the Department of Energy (DOE)-funded Genomic Encyclopedia of Bacteria and Archaea (http://www.jgi.doe.gov/programs/GEBA/index.html) and the Genome Encyclopedia of Fungi, aiming to sequence representatives for every phylogenetic node of the microbial tree of life, the sequences of more than 12 000 bacterial, 1000 fungal and dozens of protistean genomes will be obtained and stored in the international DNA databases within the next 5 yr (Kyrpides, 2009). These sequenced genomes from across the tree of life will serve as anchors for sorting through thousands of metagenomic repertoires and categorizing them.

Cataloguing the multitude

The decreasing cost and increasing throughput of DNA sequencing have prompted several research groups to embark on ribosomal RNA (rRNA) gene-sequence-based surveys of various biomes (Sogin *et al.*, 2006; Huse *et al.*, 2008), including soils (Roesch *et al.*, 2007). The maximum number of unique bacterial 16S rRNA sequences (operational taxonomic units, OTU), obtained from forest and agricultural soils, is approx. 50 000. The number of sequences needed to identify 90% of these 50 000 OTUs is lower than 1 000 000 (Roesch *et al.*, 2007; Uroz *et al.*, 2009); this could be achieved in less than a day with a single run of the current Roche 454 Titanium Genome sequencer. The efforts to monitor this unexpected diversity is, however, compounded by the fact that the rank-abundance distributions for both bacterial and fungal communities display a long tail of numerically minor, uncommon species (or OTUs) (Bent & Forney, 2008). Recent studies published in *New Phytologist* (Buée *et al.*, 2009; Jumpponen & Jones, 2009; Öpik *et al.*, 2009) demonstrated that such large-scale surveys could also be used to uncover the complexity of fungal communities in forest soils and tree phyllospheres. The number of ribosomal DNA (rDNA) sequences generated in these three studies (i.e. 309 000) is staggering, providing unique opportunities to describe fungal communities in greater detail than ever before (Hibbett *et al.*, 2009). The maximum number of predicted OTUs ranged from 1350 to 3400 per 4 g of soil, depending on the tree plantations, with a mean estimated OTU richness close to 2240 (Buée *et al.*, 2009). Despite this unexpected species diversity, most sequences of the DNA samples fell into known genera, such as *Ceratobasidium*, *Cryptococcus*, *Lactarius* and *Scleroderma*, showing that there are some dominant genera amongst the numerous organisms inhabiting soils. These genera can serve as bioindicators of environmental perturbations and are prime candidates for further functional studies. Although much higher than expected, the number of fungal OTUs detected was an order of magnitude lower than the bacterial diversity. With appropriate oligonucleotide tagging of PCR-amplified rDNA sequences, identification of the ~2000 fungal OTUs can be assessed using a fraction of a 454 pyrosequencer run (i.e. 50 000 reads). The latter seminal studies were restricted to a few sites, but given the observed responsiveness of microbial communities to shifts in environmental variables and global climate, it is of immediate concern to assess how microbiota will respond to environmental shifts at the regional and continental scales, and what impacts these shifts will have on soil health and function. Biomonitoring of microbial communities at the continental scale will rely on several thousand plots. Allowing for the necessary replicates to take into account the spatiotemporal heterogeneity, these biogeographical inventories will produce > 10 billion sequences (~4 tera (4×10^{12}) basepairs (Tbp)) per year if the viral, archaeal, bacterial, fungal and protistan communities are simultaneously surveyed throughout the European continent.

Sequencing biomes

At present, a dozen soil metagenomes have been released, which focus mostly on archaeal and bacterial species. An example of such an analysis is the functional metagenomic profiling of various biomes, including hypersaline ponds, and agricultural and forest soils (Tringe *et al.*, 2005; Dinsdale *et al.*, 2008). In the latter study, almost 15 million pyrosequencing reads from viral and microbial genomes were culled from each data set and matched to annotated genes using the metagenomics RAST server (Meyer *et al.*, 2008). This comparative gene-centric analysis showed that there are strikingly discriminatory metabolic profiles across environments, and the differences between microbiomes predicted the biogeochemical conditions of each environment. Sequencing of several million metagenomic clones is requested to obtain a representative picture of the diverse genomes present in soils. The strategy undertaken by the International Soil Metagenome Sequencing Project, the so-called TerraGenome consortium (http://terragenome.org/), is to characterize either single soil metagenomes (e.g. the Park Grass at the Rothamsted Experimental Station (Vogel *et al.*, 2009)) in detail or to explore several soils throughout landscapes and continents using environmental shotgun sequencing (Rubin, 2009). Taking into account all the ongoing projects, data sets from 4000 metagenomes will be available within the next 2 yr and they would take years or tens of years to analyse without progress in data mining and analysis. Estimates from sequencing centers involved in metagenome projects suggest that sequence data production and storage needs per annum will approach 10 Tbp of raw sequence data. This estimate does not consider the need for associated metadata, which would increase storage needs by orders of magnitude (Committee on Metagenomics, 2007). The growth of public DNA sequence data over the last two decades has been exponential, with a doubling time of about 14 months (Kyrpides, 2009). Over the next 5 yr, molecular census of microbiota structure and dynamics and cataloging gene repertoires (metagenomics) will probably be integrated to gene expression data (metatranscriptomics and metaproteomics) and soil metabolism (meta-metabolome), increasing, by several orders of magnitude, the size of the digital data sets.

The metagenome-analysis gridlock

To cope with this burst of DNA sequence data, major initiatives are needed to avoid a metagenome-analysis gridlock: the community needs to aggregate computationally intensive operations through standards and centralized coordination. To avoid wasting this data, we must switch from the traditional 'one-shot' data-mining approach to systems that are able to mine continuous, high-volume, open-ended data streams as they arrive. Scientific data collection (e.g. by earth-sensing satellites or astronomical observatories) routinely produces terabytes of data each day. Data rates of this extent have significant consequences for data mining. For example, a few months' worth of data can easily add up to billions of records, and the entire history of transactions or observations can be in the order of hundreds of billions. We can learn from colleagues working on such systems and, in order to reassure microbiologists, we will show how astrophysicists are handling the massive streams of data sets generated by sky surveys.

Unfolding the Universe

Astronomy has indeed a long tradition of systematically surveying the sky. Unable to conduct their own experiments to test theories and hypotheses, the only solution astronomers are left with is to harvest as much data as they can observe to help them to understand how stars and galaxies form and what physical rules allow the universe to unfold as we now see it. Whether aiming at finding a few rarities in the immense depths of the skies, or building up large samples of stars or galaxies in order to obtain a better statistical understanding of our Universe and how it was formed, astronomy has always been a data-driven science. Photographic plates observed at the Palomar telescope in the mid-20th century, now digitized and readily accessible (Lund & Dixon, 1973), are still used daily. Closer to us, the end of the last century has been the dawn of an era of large, systematic surveys. Tremendous leaps in our understanding of the cosmos, whether billions of light-years away, or at the doorstep of our galaxy, have been made possible by blind surveys of the night sky. The Sloan Digital Sky Survey (SDSS) (Abazajian *et al.*, 2009) is perhaps the best current example of how successful systematic observations can be. This survey builds on a rather modest 2.5-m telescope, located in New Mexico, used between 2000 and 2008 to map a quarter of the sky and obtain data on more than 200 million astronomical objects. When it was planned in the 1990s, the main driver for the survey was the mapping of the distribution of galaxies in order to better understand how visible matter distributes itself in the Universe (Gunn & Weinberg, 1995). Ten years after the first early data release, however, many fields of astronomy have been deeply modified by the SDSS and, as often, some of the most exciting discoveries have been the unexpected. To take but one spectacular example, the basic tenants of 'Galactic archaeology' (the detailed study of the surroundings of our own galaxy to find traces of how it formed and grew by the repeated absorption of smaller neighbor galaxies) have been made evident by the SDSS mapping of stars in our environs. We can now have a panoramic view (Belokurov *et al.*, 2006) of the streams of stars that are the leftovers of our cannibal galaxy's latest meals of dwarf galaxies.

Building on the SDSS success, the Large Synoptic Survey Telescope (LSST; Ivezic *et al.*, 2008) is scheduled for completion atop Chile's Cerro Pachón before 2020 and will gather, in a single night, a staggering 30 TBytes of data, which is more than the data harvest produced by the SDSS in 8 yr. Much sooner, the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS, Kaiser, 2004), already built on the Hawaiian island of Haleakala, will start operating in 2010 and will collect 1.8 PBytes of data in less than 4 yr, producing a deep mapping of three quarters of the sky. These numbers are outstanding but should not be surprising as the telescope observes one 2–3 GByte image every minute. Having each astronomer process their own useful survey data would be foolish, and, in practice, impossible. The data reduction – transforming or 'reducing' the raw observation of photons into images void of instrumental signatures, as well as scientifically useful catalogues of the properties of stars and galaxies – is therefore performed close to the telescope, by clusters of tens of computers, to distribute mainly processed data to science partners (Fig. 1). The necessary corollary is that scientists should help in building and testing an efficient data-reduction pipeline as it will impact the quality of the processed data to which they will later have access. Nonoptimal decisions have to be made during these steps so that the processing of a night's worth of images can be performed during the following day. Transient night sources, such as the rare supernovae exploding in distant galaxies, need to trigger alarms as soon as possible after they have been observed to allow for follow-up observations on other instruments. Images therefore need to be reduced quickly to avoid cluttering disk space but also because it enables the highest scientific return.

Even with the data-reduction steps performed by a dedicated team servicing the whole consortium, there remains the issue of data handling, transfer and querying. In its initial stage, Pan-STARRS regroups 13 institutes spread all around the planet, and the speed of overseas data transfers is a physical limit to what can be transferred between consortium members. Data catalogues, listing the properties of sources detected on the images, are a good compromise but they require a careful design phase to highlight the necessary data parameters and their associated quality flags that need to be distributed in order not to hinder science analyses. In the end, part of what has made SDSS so successful is the existence of an online database, which is easily accessed by every astronomer via a simple online query (http://cas.sdss.org/astrodr7/en/). Both catalogues and processed images can be queried, ensuring a versatile system that allows many science goals to be investigated, covering many astronomy fields. These questions of data management, from processing to distribution, will become so dire for the next generation of surveys that the LSST consortium has already announced a partnership with Google, whose tremendous capacity to crunch numbers and organize information will certainly strongly benefit the survey. To take the example of Pan-STARRS (Fig. 1), the reduction pipeline is written to reliably handle the daily treatment of 1.5 TBytes of data with only limited human interventions. Throughout the three-and-a-half years of the survey, it is anticipated that the Pan-STARRS consortium will face growing detection and image databases that will respectively reach 28 TBytes and 1.8 PBytes.

But, in spite of querying tools that facilitate the handling of the huge data sets produced by large, systematic sky surveys, scientists still have to learn how to efficiently mine the data in order to produce one's science. This is no small

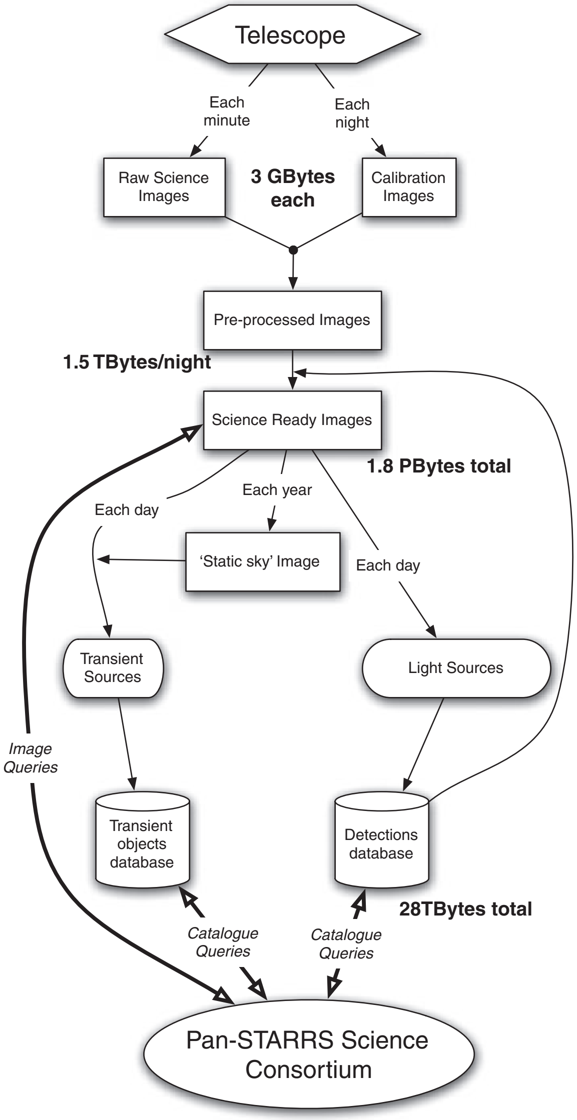

Fig. 1 The data-reduction pipeline of the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS). The flowchart shows how the data-reduction pipeline transforms photons harvested by the telescope into scientifically useful catalogues or images. Large images observed on a minute-to-minute basis are pre-processed as quickly as possible (and certainly before the next night) by using special, calibration images that are taken many times per night. The next step consists of tying the images on global reference frames in order to know, for instance, the absolute position of each object detected on an image. Once the images have been entirely calibrated, astronomical sources are detected and their properties are stored in a large database. In parallel, the Pan-STARRS pipeline will produce a 'static sky' image to detect, more easily, objects with fluctuating positions or luminosities. This image of the whole sky represents the current best knowledge of what the night sky looks like. It is subtracted from each observed image to allow the detection of transient sources: fast moving sources, such as close-by asteroids that will blink in and out of the difference image; or fluctuating sources, such as a remote galaxy that has its luminosity increasing sharply from the explosion of one of its stars into a supernova.

endeavour and requires a rethinking of one's path for producing exciting results. Where one would, years ago, study the properties of a few tens or hundreds of 'objects', whether they be galaxies or genes, one needs to turn to approaches that rely heavily on statistical analyses to efficiently milk thousands or millions of catalogue entries. Whether one is interested in the bulk properties of a population (e.g. the distribution of surrounding stars in the Milky Way to constrain its structure, or ectomycorrhizal fungi in forest soils), or focusing on identifying individual outliers, which are hard to find but likely to produce a high scientific return (e.g. the very rare galaxies of the first ages of the Universe probing the local conditions of these remote times, or the very rare bacterium that will provide increased resilience to the aboveground plant community in stressful times), both approaches require sound statistical foundations.

Cyberinfrastructure for Microbial Metagenomics

The challenge facing microbiologists playing with (meta)genomes cannot be achieved by the efforts of individual researchers, but requires the establishment of effective national and international collaborations as did the (astro)physicists to rapidly set up the most efficient, robust informatic pipelines, but also to successfully lobby for access to the largest supercomputer grids. Two large projects have recently been initiated to build an infrastructure for metagenomic sequences and associated metadata: the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) project (http://camera.calit2.net/) (Seshadri *et al.*, 2007), and the Integrated Microbial Genomes (IMG) system (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) (Markowitz *et al.*) at the DOE's Joint Genome Institute (Walnut Creek, CA). The objective of the CAMERA project is to provide cyberinfrastructure tools and resources, and bioinformatics expertise to enable the community to use metagenomic data. CAMERA and MG-RAST provide a public BLAST search and annotation service, enabling large-scale comparisons to be made against a predefined set of databases in a way similar to the Pan-STARRS pipeline (Fig. 1). Visualization of the taxonomic composition of metagenomic samples after BLAST, and comparison of taxonomic composition between samples, can be carried out using tools such as MEGAN (Huson *et al.* 2007). However, these programs only perform individual steps of metagenomic analysis and we lack a truly integrated solution in which analyses can be easily concatenated, converted to workflows, shared among colleagues and published in a readily reproducible form. Such novel comprehensive pipelines for phylogenetic profiling of metagenomic samples, which includes all steps, from processing and quality control of data generated by next-generation sequencing technologies, to statistical analyses and data visualization, are in development (Pond *et al.* 2009).

The skies are clearing for microbiologists as discussions at the recent International Conference on Systems for Intelligent Molecular Biology led community members to form the M5 (metagenomics, metadata, meta-analysis, multi-scale-models and meta-infrastructure) Consortium under the roof of the Genomics Standards Consortium to devise a solution to the forecasted digital gridlock. Their proposed 'M5 Platform' – to be announced later this year – deserves the support of our community, funding agencies and those who hold the keys to the high-performance computing centers. If the metagenome-analysis gridlock is unlocked, we will 'know as much about the soil underfoot as we do about the movement of celestial bodies' in a decade. The blind survey of the streams of microbial sequences will undoubtly facilitate the understanding of the mechanisms ruling the subterranean communities and bring exciting, unexpected discoveries. With the advent of these new tools and techniques, the possibility of integration across a wide range of disciplines, from genomics to molecular ecology and field ecology, is becoming a reality that is much encouraged by *New Phytologist*.

Nicolas F. Martin

Max-Planck-Institut für Astronomie, Königstuhl 17,
D-69117 Heidelberg, Germany

Francis Martin

Interaction Section Editor, *New Phytologist*
(email fmartin@nancy.inra.fr)

Acknowledgements

Research in FM's lab is funded by grants from the European Network of Excellence EVOLTREE, the EU ENERGYPOPLAR programme and the Région Lorraine Council (Project FORBOIS). NFM is supported by the Max-Planck-Institut für Astronomie. We would like to thank Dr Holly Slater for careful reading of the mansucript.

References

Abazajian KN, Adelman-McCarthy JK, Agüeros MA, Allam SS, Allende Prieto C, An D, Anderson KSJ, Anderson SF, Annis J, Bahcall NA *et al.* 2009. The seventh data release of the Sloan Digital Sky Survey. *The Astrophysical Journal. Supplement Series* **182**: 543–558.

Belokurov V, Zucker DB, Evans NW, Gilmore G, Vidrih S, Bramich DM, Newberg HJ, Wyse RFG, Irwin MJ, Fellhauer M *et al.* 2006. The field of streams: Sagittarius and its siblings. *The Astrophysical Journal* **642***: L137–L140.*

Bent SJ, Forney LJ. 2008. The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *The ISME Journal* **2**: 689–695.

Buée B, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.

Committee on Metagenomics (2007). *The new science of metagenomics – revealing the secrets of our microbial planet.* Washington, DC, USA: Committee on Metagenomics: Challenges and Functional Applications. National Research Council of The National Academies. The National Academies Press.

Curtis TP, Sloan WT. 2006. Exploring microbial diversity – a vast below. *Science* **309**: 1331–1333.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L *et al.* 2008. Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Gans J, Wolinsky M, Dunbar J. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.

Gunn J, Weinberg D (1995). The Sloan Digital Sky Survey. Wide field spectroscopy and the distant universe. In: Maddox SJ, Aragón-Salamanca A, eds. *Proceedings of the 35th Herstmonceux Conference, held July 4–8, 1994, Cambridge, UK.* Singapore: World Scientific, 1995, 3.

Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669–685.

Hibbett DS, Ohman A, Kirk PM. 2009. Fungal ecology catches fire. *New Phytologist* **184**: 279–282.

Hugenholtz P, Tyson GW. 2008. Microbiology: metagenomics. *Nature* **455**: 481–483.

Huse SM, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genetics* **4**: e1000255.

Huson DH, Auch AF, Qi Ji, Schuster SC. 2007. MEGAN Analysis of metagenomic data. *Genome Research* **17**: 377–386.

Ivezic Z, Tyson JA, Allsman R, Andrew J, Angel R. 2008. LSST: from science drivers to reference design and anticipated data products. eprint arXiv:0805.2366 http://www.lsst.org/lsst/science/overview.

Jumpponen A, Jones KL. 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* **184**: 438–448.

Kahvejian A, Quackenbush J, Thompson JF. 2008. What would you do if you could sequence everything? *Nature Biotechnology* **26**: 1125–1133.

Kaiser N (2004). *Pan-STARRS: a wide-field optical survey telescope array.* In: Oschmann JM Jr, ed. *Proceedings – Society of Photo-Optical Instrumentation Engineers* **5489**: 11–22.

Kyrpides N. 2009. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nature Biotechnology* **27**: 627–632.

Lund JM, Dixon RS. 1973. A user's guide to the Palomar Sky survey. *Publications of the Astronomical Society of the Pacific* **85**: 230.

Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I *et al.* 2006. The integrated microbial genomes (IMG) system. *Nucleic Acids Research* **34**: D344–D348.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A *et al.* 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M. 2009. Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist* **184**: 424–437.

Pond SK, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J, Nekrutenko A, the Galaxy Team. 2009. Windshield splatter analysis with the Galaxy metagenomic pipeline. *Genome Research* **11**: 2144–2153.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, Daroub SH, Camargo FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal* 1: 283–290.

Rubin E (2009). JGI overview now and the future. http://www.science.doe.gov/ober/berac/Rubin09-09.pdf.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K *et al.* 2007. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5: e77.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. 2007. CAMERA: a community resource for metagenomics. *PLoS Biology* 5: e75.

Singh BK, Campbell CD, Sorenson SJ, Zhou J. 2009. Soil genomics. *Nature Reviews. Microbiology* 7: 756.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences, USA* **103**: 12115–12120.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al.* 2005. Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Uroz S, Buée M, Murat C, Frey-Klett P, Martin F (2009). Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environmental Microbiology Reports*, in press.

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R & Philippot L. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews. Microbiology* 7: 252.

Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences, USA* **95**: 6578–6583.

Key words: (meta)genomics, astrophysics, biomes, DNA sequencing, galactic archaeology, massive data sets, molecular (microbial) ecology, super computer grids.

Commentary

Plant functional traits – linkages among stem anatomy, plant performance and life history

Plant ecology in recent years has focused on functional traits (Westoby & Wright, 2006; Swenson & Enquist, 2007; Chave *et al.*, 2009; Cornwell & Ackerly, 2009) – phenotypic attributes that influence performance in a given environmental setting. As a starting point for cross-species comparisons, the traits selected have been easy to measure (Cornelissen *et al.*, 2003). Examples include seed size, leaf size, maximum plant height, leaf mass per area, leaf area per stem area and wood density. But, even with large sample sizes across many species, the interpretation of results can be complex; most of these traits reflect multiple aspects of plant function. Wood density has attracted considerable interest. It has been related to growth rate, water storage, mechanical strength, efficiency and safety of hydraulic transport, and resistance to herbivory (Jacobsen *et al.*, 2007; Sperry *et al.*, 2008; Chave *et al.*, 2009; Onoda *et al.*, this issue, pp. 493–501). However, with so many potential competing functions influencing how densely a stem is built, it becomes hard to determine the utility of having denser or lighter wood in a particular setting. One way to make headway is to examine the harder-to-measure anatomical traits that contribute to lighter or denser wood and apply our understanding of the functional roles that these different anatomical tissues play. In the current issue of *New Phytologist*, Poorter *et al.*, (pp. 481–492) have taken just such an approach. They quantified several of these harder-to-measure anatomical traits and examined their relationship to both wood density and various aspects of plant performance and life history for 42 abundant tree species in the rainforests of La Chonta, Bolivia.

> *'... while plants have relatively little room to adjust the percentage of stem devoted to vessels, they seem to have ample leeway in deciding how this space is divided up.'*

What drives variation in wood density across species?

Poorter *et al.* measured the stem cross-sections from each of the species as the fractions composed of vessel, fibre and parenchyma (Fig. 1). These three tissues have presumably evolved to meet the challenges faced by woody angiosperm stems. Vessels are the main conducting cells in angiosperms, with flow rates per vessel increasing with the cross-sectional area of the lumen (Tyree & Zimmermann, 2002). Fibres are the main support structures, and parenchyma serves for both storage and transport of resources between xylem and phloem (Martínez-Cabrera *et al.*, 2009). Depending on selective pressures, evolution may

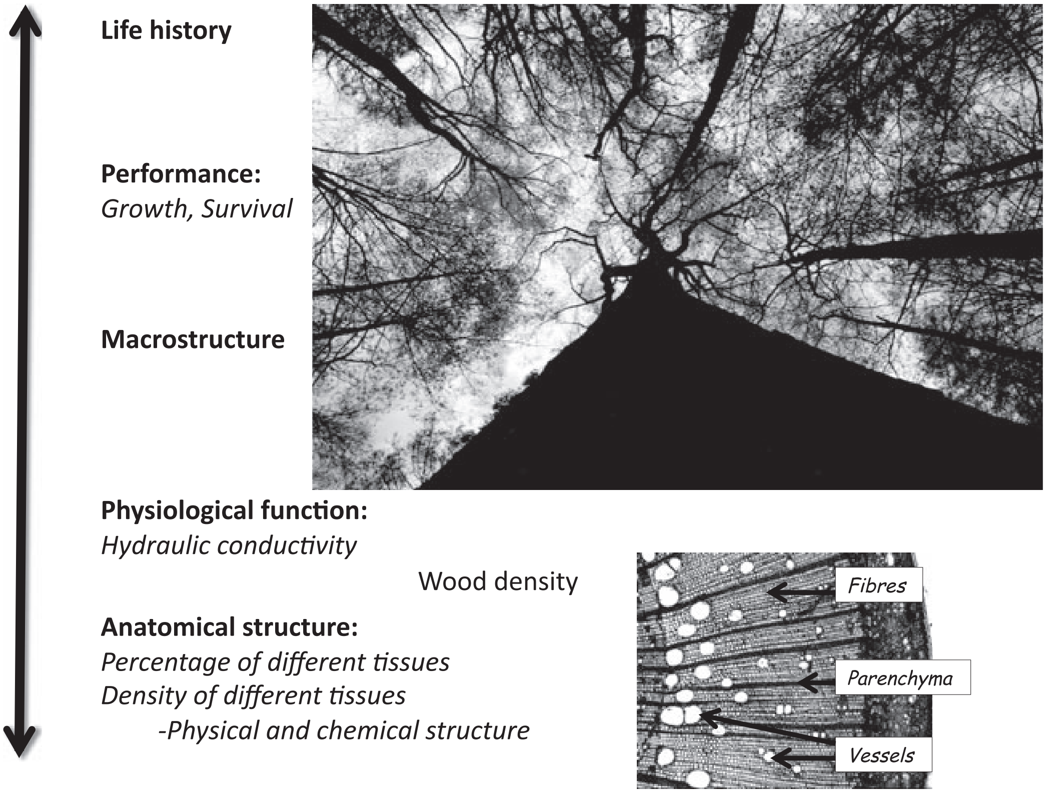

Fig. 1 Examining trait functionality. Representation of linkages between plant functional traits at the anatomical level to physiological function, macrostructure, plant performance (growth and survival) and life history, as examined in the study of Poorter *et al.* (this issue; pp. 481–492).

alter the relative allocation to these different tissue types, or, alternatively, the individual densities (by shifts in chemical or structural composition) of the different tissues. Both types of change are reflected in an aggregate measure such as wood density. Partitioning variation in wood density into these components, however, should enhance our understanding of the functional significance of lighter or denser wood.

One exciting result arising from the study of Poorter *et al.* is their finding that variation in the density of the different tissue types, rather than the relative amount of each tissue, appears to underpin variation in wood density across these species (see also Zanne *et al.*, 2010). As vessels take up volume but have limited mass (because of their large lumens), a relationship between wood density and fraction of cross-sectional area that is vessel might be expected. Similarly, because fibres are heavy as a result of thick walls, we might expect a relationship between wood density and the fraction of the cross-sectional area that is fibre. As it turned out, neither relationship was particularly strong. These results point to variation in tissue composition, rather than to relative amounts of vessel, parenchyma and fibre, as being the primary determinants of wood density. Supporting this idea, wood density has been related to fibre wall to lumen ratios, percentage allocation to fibre wall and fibre lumen size (Jacobsen *et al.*, 2007; Martínez-Cabrera *et al.*, 2009).

The lack of a relationship between total percentage vessel area and density is particularly significant, because it brings into question the suggested links between wood density and hydraulic conductivity. As well as making wood lighter, greater vessel area should improve conductivity, leading to predictions for relationships between wood density and conductivity (Meinzer, 2003). But, as Poorter *et al.* show, greater allocation to vessels does not necessarily lead to lighter wood, so links between conductivity and density should not be taken as a given.

Cross-species indicators of vascular strategy

The limited variation in percentage vessels (3–23%), observed by Poorter *et al.*, contrasts with the nearly 500-fold variation observed in average vessel size (A) and number of vessels per unit area (or density of vessels, N) (Fig. 2). These two variables show a strong, negative correlation (see also Preston *et al.*, 2006; Sperry *et al.*, 2008; Zanne *et al.*, 2010), indicating coordinated shifts in the mixture of vessel sizes and numbers across species. So, while plants have relatively little room to adjust the percentage of stem devoted to vessels, they seem to have ample leeway in deciding how this space is divided up. Furthermore, because potential conductivity increases to the square of A but only to the first power of N (Tyree & Zimmermann, 2002), altering the

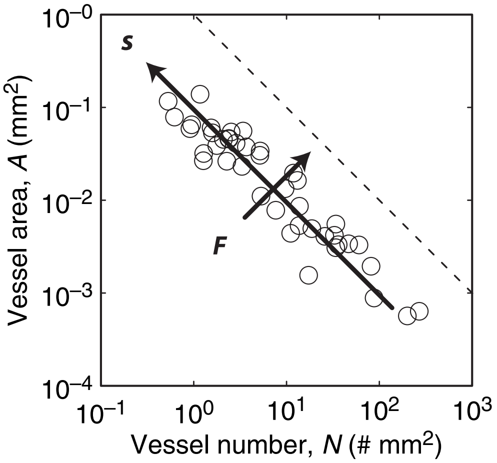

Fig. 2 Variation in vascular design of 42 abundant tree species in La Chonta, Bolivia, as reported by Poorter *et al.* (this issue; pp. 481–492). Although most studies report average vessel size (*A*) and vessel number (*N*), vascular strategy can also be represented as vessel size : number ratio (*S*) and vessel fraction (*F*). The dashed line shows the packing limit for vessel fraction, given by *F* = 1.

size–number mixture can have large influences on potential conductivity.

Poorter *et al.* – like other studies before them – adopt vessel size and number as the primary indicators of vascular strategy. We agree that vessel size and number are critical aspects of plant strategy, but note that two difficulties arise when using *A* and *N* in comparative studies. The first is that they are tightly correlated, and so do not provide independent information. The second is that neither *A* nor *N* distinguishes between the different ways in which species can adjust rates of water supply through sapwood. Cross-species variation in *A* or *N* could indicate changes to the total area of conducting tissue (vessel fraction), the mixture of vessel sizes and numbers, or both. These difficulties are overcome if one instead uses vessel fraction ($F = AN$, mm^2) and the vessel size : number ratio (S; $=A/N$, mm^4) as indicators of vascular strategy (Fig. 2).

We recently analyzed global patterns in *S* and *F* for over 2000 woody angiosperm species distributed worldwide (Zanne *et al.*, 2010) and showed that the bulk of variation in *A* and *N* equates to variation in the size : number ratio *S* (95%), leaving only 5% accounted for by vessel fraction *F*. This decomposition indicates that shifts in the size : number ratio are therefore also responsible for most of the variation in potential conductivity observed across species. Analysis of the data of Poorter *et al.* recovered the same results (Fig. 2), demonstrating that such global patterns may also be observed within local studies. In fact, further analysis of the data of Poorter *et al.* reveal that all of the reported associations between vessel anatomy and plant performance or life history (see below) are related to shifts in the mixture of vessel sizes and numbers, and not to shifts in vessel fraction. Additionally, their data indicate that most of the cross-species variation in *S* and *F* observed worldwide can be found within a single community.

Does trait variation lead to variation in plant performance?

A consistent and outstanding feature of the publications from Poorter and colleagues in Bolivia and the Netherlands is their ongoing effort to link functional traits with plant performance (growth and survival) and life-history strategy (maximum adult height and sapling crown exposure). As with other studies of theirs (e.g. van Gelder *et al.*, 2006; Poorter & Bongers, 2006), performance measures have been quantified across numerous individuals, allowing them to take the significant step of testing how functional these functional traits really are. Confirming previous findings, wood density was strongly related to both growth and survival (Chave *et al.*, 2009). However, vessel size and number also had strong influences on growth rates. As relationships with the vessel fraction *F* were not significant, these must have been mediated through changes in the size : number ratio, *S*. The study of Poorter *et al.* thus supports two independent influences in stem design leading to faster growth rates, namely more efficient construction of stems (via shifts to low wood density) and higher hydraulic conductivity, enabling faster photosynthetic rates (via shifts to large *S*). To our knowledge, this is the first study to show that vessel traits related to high hydraulic conductivity form part of a spectrum of traits leading to fast growth; other traits that show this linkage include leaf mass per area, leaf nitrogen, wood density and stem allometry (Poorter & Bongers, 2006; Westoby & Wright, 2006; Chave *et al.*, 2009).

A widely held view in ecology is that plants differentiate along successional niche and/or growth strategy axes, and that this differentiation has both generated and maintained diversity in tropical forests. However, it has been unclear how wood structure contributes to this differentiation. Thanks to detailed studies, such as that of Poorter *et al.*, our understanding of these issues is improving; challenges, however, remain. The costs of fast growth, via high *S* or low wood density, are still poorly documented. The vessel-size number axis is presumed to represent a trade-off between conductivity and hydraulic safety (driven by embolism avoidance). But, does embolism avoidance limit success in shaded environments, and, if so, how? Similarly, low wood density is known to decrease the structural integrity of stems (van Gelder *et al.*, 2006; Jacobsen *et al.*, 2007; Onoda *et al.*, 2010), but it remains unclear how this limits fast-growing, light-wooded species from recruiting in low light. These are important questions for future studies that seek to bridge the gap between functional traits, plant performance and life-history strategy (Fig. 1).

Amy E. Zanne[1]* and Daniel S. Falster[2]

[1]Department of Biology, University of Missouri, St Louis, St Louis, MO 63108, USA; [2]Biological Sciences, Macquarie University, NSW 2109, Australia
(*Author for correspondence: tel +1 314 516 6672; email aezanne@gmail.com)

References

Chave J, Coomes D, Jansen S, Lewis SL, Swenson NG, Zanne AE. 2009. Towards a worldwide wood economics spectrum. *Ecology Letters* **12**: 351–366.

Cornelissen JHC, Lavorel S, Garnier E, Díaz S, Buchmann N, Gurvich DE, Reich PB, ter Steege H, Morgan HD, van der Heijdetn MGA *et al.* 2003. Handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Australian Journal of Botany* **51**: 335–380.

Cornwell WK, Ackerly DD. 2009. Community assembly and shifts in the distribution of functional trait values across an environmental gradient in coastal California. *Ecological Monographs* **79**: 109–126.

van Gelder HA, Poorter L, Sterck FJ. 2006. Wood mechanics, allometry, and life-history variation in a tropical rain forest tree community. *New Phytologist* **171**: 367–378.

Jacobsen AL, Agenbag L, Esler KJ, Pratt RB, Ewers FW, Davis SD. 2007. Xylem density, biomechanics and anatomical traits correlate with water stress in 17 evergreen shrub species of the Mediterranean-type climate region of South Africa. *Journal of Ecology* **95**: 171–183.

Martínez-Cabrera HI, Jones CS, Espino S, Schenk HJ. 2009. Wood anatomy and wood density in shrubs: responses to varying aridity along transcontinental transects. *American Journal of Botany* **96**: 1388–1398.

Meinzer FC. 2003. Functional convergence in plant responses to the environment. *Oecologia* **134**: 1–11.

Onoda Y, Richards A, Westoby M. 2010. The relationship between stem biomechanics and wood density is modified by rainfall in 32 Australian woody plant species. *New Phytologist* **185**: 493–501.

Poorter L, Bongers FJJM. 2006. Leaf traits are good predictors of plant performance across 53 rain forest species. *Ecology* **87**: 1733–1743.

Poorter L, McDonald I, Alarcón A, Fichtler E, Licona J-C, Peña-Claros M, Sterck F, Villegas Z, Sass-Klaassen U. 2010. The importance of wood traits and hydraulic conductance for the performance and life history strategies of 42 rainforest tree species. *New Phytologist* **185**: 481–492.

Preston KA, Cornwell WK, DeNoyer JL. 2006. Wood density and vessel traits as distinct correlates of ecological strategy in 51 California coast range angiosperms. *New Phytologist* **170**: 807–818.

Sperry JS, Meinzer FC, McCulloh KA. 2008. Safety and efficiency conflicts in hydraulic architecture: scaling from tissues to trees. *Plant, Cell & Environment* **31**: 632–645.

Swenson NG, Enquist BJ. 2007. Ecological and evolutionary determinants of a key plant functional trait: wood density and its community-wide variation across latitude and elevation. *American Journal of Botany* **94**: 451–459.

Tyree MT, Zimmermann MH. 2002. *Xylem structure and the ascent of sap*. Berlin, Germany: Springer.

Westoby M, Wright IJ. 2006. Land-plant ecology on the basis of functional traits. *Trends in Ecology & Evolution* **21**: 261–268.

Zanne AE, Westoby M, Falster DS, Ackerly DD, Loarie SR, Arnold SEJ, Coomes DA. 2010. Angiosperm wood structure: global patterns in vessel anatomy and its relationship to wood density and potential conductivity. *American Journal of Botany*. In press.

Key words: angiosperms, cross-species comparisons, functional traits, life history, plant performance, stem anatomy, wood density.

General latitudinal gradient of biodiversity is reversed in ectomycorrhizal fungi

Tropical rainforests support a tremendous biodiversity of animals and plants, but because of difficulties in observation and identification, little is known about the richness patterns of microbes, including fungi. The general latitudinal gradient of diversity (LGD) demonstrates that nearly all terrestrial and marine macroorganisms studied so far have peak richness at low latitudes (Hillebrand, 2004). Indeed, recent research in the Americas revealed that neotropical rainforests comprise the highest diversity of endophytic fungi. Plant taxa harbor different endophyte communities and therefore fungal diversity increases with host diversity (reviewed in Arnold, 2008). Similar patterns may enhance the diversity of ectomycorrhizal fungi (EcMF) in the temperate zone (Bruns *et al.*, 2002; Ishida *et al.*, 2007; Tedersoo *et al.*, 2008), but the relevance of this and other potential factors remain unknown in tropical ecosystems because of a lack of published studies. In this issue of *New Phytologist*, Peay *et al.* (pp. 529–542) address the biotic and abiotic factors driving the composition of an EcMF community in a Bornean rainforest that comprises one of the highest plant diversities in the world.

> *'... many of the proposed causal mechanisms that have been developed to explain the general LGD pattern may be inapplicable to EcMF or need substantial modification.'*

Do tropical and temperate EcMF communities differ? Extensive fruit-body surveys have revealed that tropical and temperate forests share many common EcMF lineages, such as the /boletus, /cantharellus, /clavulina, /russula–lactarius, /tomentella–thelephora, etc. (reviewed in Tedersoo

Table 1 Site descriptions of four temperate and four tropical forests used

Site	Geocode	Vegetation type	Site size (ha)	Host lineages	Reference
Huizteco, Mexico	18°36′N 99°36′W	Subtropical montane cloud forest	3	Fagales	Morris *et al.* (2009)
Mt Field, Tasmania	42°41′S 145°42′W	Wet sclerophyll forest	20	Leptospermoideae, Fagales, Pomaderreae	Tedersoo *et al.* (2008)
Sierra Foothill, California, USA	39°17′N 121°17′W	Mediterranean woodland	50	Fagales, Pinaceae	Morris *et al.* (2008); Smith *et al.* (2009)
Tagamõisa, Estonia	58°45′N 27°00′E	Temperate woodland	10	Fagales, Pinaceae, Salicaceae, *Tilia*	Tedersoo *et al.* (2006)
Bukit Bangkirai, Indonesia	01°01′S 116°52′E	Tropical rainforest	30	Dipterocarpaceae Fagales	K. Nara *et al.* (unpublished)
Lambir Hills, Malaysia	04°20′N 113°50′E	Tropical rainforest	18	Dipterocarpaceae, Fagales, Leptospermoideae[1], Detariae[1]	Peay *et al.* (2010) (this issue)
Monts de Cristal, Gabon	00°37′N 10°25′E	Tropical rainforest	20	Amhersteae, Dipterocarpaceae, *Uapaca*	L. Tedersoo *et al.* (unpublished)
Yasuni, Ecuador	00°37′S 76°24′W	Tropical rainforest	30	Pisoniae, Coccolobae	Tedersoo *et al.* (2010b)

[1] It remains unknown whether the root tips of these taxa were actually sampled.

et al., 2010a). However, EcMF fruit-bodies do not provide useful data for quantitative comparisons of the local EcMF diversity among studies, because of their strongly seasonal and yearly unpredictable production and ephemeral habit. Several common EcMF lineages, such as the /tomentella–thelephora, /piloderma and /hysterangium, produce either resupinate fruit-bodies on the underside of debris or hypogeous sporocarps that remain unnoticed unless specifically searched for. Moreover, many ascomycete lineages and some basidiomycete lineages probably lack fruit-bodies. The development of molecular techniques has facilitated the identification of EcMF from root tips and mycelia that can be found in soil throughout the year. Rapidly advancing sequencing and microarray technologies enable quantitative analyses of EcMF diversity and ecology based on numerous independent soil or root samples. Peay *et al.* collected 101 soil samples from seven 400-m^2 plots at Lambir Hills, Borneo, and recovered 105 EcMF species, most of which belong to the /russula–lactarius, /boletus and /tomentella–thelephora lineages. These data are valuable for shedding light on EcMF diversity and community composition in tropical habitats, but do not represent all tropical forests that differ in historical biogeography and composition of ectomycorrhizal host lineages. To address differences in EcMF diversity and community structure between temperate and tropical rainforests, we supplemented the data from Lambir Hills with three data sets from different tropical regions (Africa, South America and South East Asia) and compared them with the results of temperate studies that employed generally similar methods for sampling and identification (Table 1).

Across these studies, the most species-rich lineages were /thelephora–tomentella and /russula–lactarius, followed by /cortinarius, /sebacina, /clavulina, /boletus and /inocybe (Supporting Information Fig. S1). Surprisingly, the relative richness of only two EcMF lineages differed significantly between tropical and temperate habitats – /russula–lactarius (*t*-test: $t = 18.8$; $P = 0.005$) and /inocybe ($t = 8.4$; $P = 0.027$), which were relatively more diverse in tropical and temperate forests, respectively. However, the statistical significance collapsed when adjusted using the Bonferroni correction for multiple tests.

In contrast to the compositional similarity of EcMF, the diversity of EcMF in tropical sites was lower than that in temperate sites, as revealed by the rarefied species richness (Fig. 1a; *t*-test: $t = 61.3$; $P < 0.001$) and minimal richness estimators – Jackknife 2, Chao2 and ICE (Fig. S2). This pattern in EcMF strongly contradicts the negative LGD (Hillebrand, 2004). Thus, many of the proposed causal mechanisms, such as the mid-domain, geographical area and climatic stability effects that have been developed to explain the general LGD pattern (Lomolino *et al.*, 2006), may be inapplicable to EcMF or need substantial modification.

What, then, causes the lower EcMF richness in tropical ecosystems? Although the information available is scant, we propose three possible mechanisms. First, historical and biogeographical effects may partly explain the differences in EcMF richness pattern observed between tropical and temperate ecosystems. Many EcMF lineages occur in both tropical and temperate biomes. While several temperate lineages are notably absent from tropical forests, no strictly tropical lineages are known (Tedersoo *et al.*, 2010a). The EcMF community data support this observation: temperate forests harbor more EcMF lineages than tropical forests (Fig. 1b; *t*-test: $t = 345.4$; $P < 0.001$). The strictly temperate EcMF lineages probably evolved at higher latitudes with the Pinaceae hosts, but may be inferior competitors in tropical conditions. However, this can only partly explain the observed pattern, because these temperate lineages are relatively species-poor and usually form a minor component in temper-

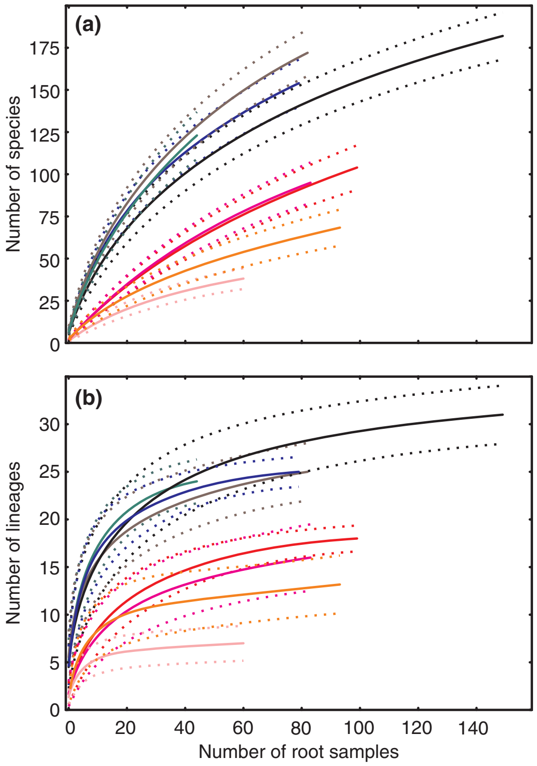

Fig. 1 Rarefied accumulation curves of (a) species and (b) lineages of ectomycorrhizal fungi (EcMF) in four temperate and four tropical sites. Dotted lines indicate 95% CI. Brown, Tagamõisa, Estonia; green, Mt Field, Australia; blue, Huizteco, Mexico; black, UC Sierra Foothill, CA, USA; purple, Lambir Hills, Malaysia; red, Monts de Cristal, Gabon; orange, Bukit Bangkirai, Indonesia; pink, Yasuni, Ecuador. The minimal richness estimates were calculated based on original data sets by use of a computer program ESTIMATE S (Colwell, 2006; ESTIMATE S: statistical estimation of species richness and shared species from samples. Version 8. Persistent URL \lt purl.oclc.org/ estimates \gt) and 1000 replicates.

ate forests (except for the /amphinema–tylospora and strictly Pinaceae-specific /suillus lineages).

Second, when both soil and roots are regarded as habitats for EcMF, the lower diversity and abundance of these habitats may account for the lower EcMF diversity in the tropics. The roots of host plants are an obligatory energy source for all EcMF. Therefore, the co-existence of different hosts enhances habitat diversity. With certain exceptions (Morris *et al.*, 2008), host preference is rarely evident at the host species level, but is probably more important at higher taxonomic levels, that is, from genus to phylum (Ishida *et al.*, 2007). While tropical forests are often dominated by a single ectomycorrhizal host lineage, such as the Dipterocarpaceae or certain groups of Fabaceae, temperate habitats are often composed of multiple codominant host lineages, for example, Fagales, Pinaceae and Salicaceae. Although many different dipterocarp or leguminous host species co-exist in a tropical site, their phylogenetic habitat difference is lower compared with temperate forests. Similarly, the soil habitat is less diverse in the tropics than in temperate forests. In tropical forests, the soil profile is usually poorly differentiated and has thin organic and litter layers as a result of the rapid consumption of organic matter by mesofauna and the fast decomposition rates that occur in constantly warm and humid conditions. Many temperate EcMF species display niche differentiation by soil horizons (e.g. Lindahl *et al.*, 2007) and EcMF communities accumulate more species in better developed soils (Nara *et al.*, 2003).

Third, resource availability and fragmentation may explain the observed pattern of EcMF richness. Many temperate forests are exclusively dominated by suitable host trees that may account for 100% in basal area or the number of stems. Conversely, EcMF hosts usually contribute < 75% to the basal area in South East Asia and in monodominant patches of other tropical regions. Tropical EcM hosts are most often distributed sparsely, forming small isolated host islands in non-EcM vegetation. Fragmentation and availability of fewer EcM roots may reduce the population size of EcMF and eventually result in fewer co-existing EcMF species (Peay *et al.*, 2007; Tedersoo *et al.*, 2010b)

The comparison of temperate and tropical forests demonstrates that the EcMF richness pattern is an exception to the general LGD. The underlying mechanisms of the EcMF richness pattern may include historical–biogeographical factors, habitat diversity and resource availability–fragmentation. Accumulating EcMF data from seasonal tropical and subarctic ecosystems will probably improve our understanding of these causal mechanisms and the entire shape of the latitudinal gradient of EcMF diversity. In addition to high-quality DNA sequence data, collection of various environmental and biological data for each study site is essential to develop a general model of EcMF diversity and community composition on a global scale (Lilleskov & Parrent, 2007).

Leho Tedersoo[1]* and Kazuhide Nara[2]

[1]Institute of Ecology and Earth Sciences and Natural
History Museum of Tartu University; 40 Lai, 51005 Tartu,
Estonia; [2]Asian Natural Environmental Science Center,
The University of Tokyo, Midori-cho 1-1-8, Nishi-Tokyo,
Tokyo 188-0002, Japan
(*Author for correspondence: tel +372 56 654 986;
email leho.tedersoo@ut.ee)

References

Arnold AE. 2008. Endophytic fungi: hidden components of tropical community ecology. In: Carson WP, Schnitzer SA, eds. *Tropical forest community ecology.* Oxford, UK: Wiley-Blackwell, 254–271.

Bruns TD, Bidartondo MI, Taylor DL. 2002. Host specificity in ectomycorrhizal communities: what do the exceptions tell us? *Integrative and Comparative Biology* 42: 352–359.

Colwell RK. 2006. *EstimateS: statistical estimation of species richness and shared species from samples. Version 8*. Available at: http://purl.oclc.org/estimates. Last accessed 30.11.2009.

Hillebrand H. 2004. On the generality of the latitudinal diversity gradient. *American Naturalist* 163: 192–211.

Ishida TA, Nara K, Hogetsu T. 2007. Host effects on ectomycorrhizal fungal communities: insight from eight host species in mixed conifer–broadleaf forests. *New Phytologist* 174: 430–440.

Lilleskov EA, Parrent JL. 2007. Can we develop general predictive models of mycorrhizal fungal community–environment relationships? *New Phytologist* 174: 250–256.

Lindahl B, Ihrmark K, Boberg J, Trumbore SE, Högberg P, Stenlid J, Finlay RD. 2007. Spatial separation of litter decomposition and mycorrhizal nutrient uptake in a boreal forest. *New Phytologist* 173: 611–620.

Lomolino MV, Riddle BR, Brown JH. 2006. *Biogeography*, 3rd edn. Sunderland, MA, USA: Sinauer Associates, Inc.

Morris MH, Smith ME, Rizzo DM, Rejmanek M, Bledsoe CS. 2008. Contrasting ectomycorrhizal fungal communites on the roots of co-occuring oaks (*Quercus* spp.) in a California woodland. *New Phytologist* 178: 167–176.

Morris MH, Perez-Perez MA, Smith ME, Bledsoe CS. 2009. Influence of host species on ectomycorrhizal communities associated with two co-occurring oaks (*Quercus* spp.) in a tropical cloud forest. *FEMS Microbiology Ecology* 69: 274–287.

Nara K, Nakaya H, Wu B, Zhou Z, Hogetsu T. 2003. Underground primary succession of ectomycorrhizal fungi in a volcanic desert on Mount Fuji. *New Phytologist* 159: 743–756.

Peay KG, Bruns TD, Kennedy PG, Bergemann SE, Garbelotto M. 2007. A strong species-area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecology Letters* 10: 470–480.

Peay KG, Kennedy PG, Davies SJ, Tan S, Bruns TD. 2010. Potential link between plant and fungal distribuions in a dipterocarp rainforest: community and phylogenetic structure of tropical ectomycorrhizal fungi across a plant and soil ecotone. *New Phytologist* 185: 529–542.

Smith ME, Douhan GW, Fremier AK, Rizzo DM. 2009. Are true multi-host fungi the exception or the rule? Dominant ectomycorrhizal fungi on *Pinus sabiniana* differ from those on co-occurring *Quercus* species *New Phytologist* 182: 295–299.

Tedersoo L, Suvi T, Larsson E, Kõljalg U. 2006. Diversity and community structure of ectomycorrhizal fungi in a wooded meadow. *Mycological Research* 110: 734–748.

Tedersoo L, Jairus T, Horton BM, Abarenkov K, Suvi T, Saar I, Kõljalg U. 2008. Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. *New Phytologist* 180: 479–490.

Tedersoo L, May TW, Smith ME. 2010a. Ectomycorrhizal lifestyle in fungi: global diversity, distribution, and evolution of phylogenetic lineages. *Mycorrhiza* doi: 10.1007/s00572-009-0274-x.

Tedersoo L, Sadam A, Zambrano M, Valencia R, Bahram M. 2010b. Low diversity and high host preference of ectomycorrhizal fungi in Western Amazonia, a neotropical biodiversity hotspot. *The ISME Journal*, doi:10.1038/ismej.2009.131.

Key words: biogeography, community structure, latitudinal diversity gradient, soil microbes, temperate forest, tropical rainforest.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Relative contribution of 13 most common lineages of EcMF to the local species richness in tropical and temperate forests.

Fig. S2 Rarefied accumulation curves of minimal species richness estimates for the eight study sites in tropical and temperate forests.

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