

## Structure, Fluctuation and Magnitude of a Natural Grassland Soil Metagenome

Tom O. Delmont<sup>1</sup>, Emmanuel Prestat<sup>1</sup>, Kevin P. Keegan<sup>2</sup>, Michael Faubladier<sup>3</sup>, Patrick Robe<sup>4</sup>, Ian M. Clark<sup>5</sup>, Eric Pelletier<sup>6,7,8</sup>, Penny R. Hirsch<sup>5</sup>, Folker Meyer<sup>2</sup>, Jack A Gilbert<sup>2,9</sup>, Denis Le Paslier<sup>6,7,8</sup>, Pascal Simonet<sup>1</sup> and Timothy M. Vogel<sup>1\*</sup>

5 <sup>1</sup> Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France.

<sup>2</sup> Institute of Genomic and Systems Biology, Argonne National Laboratory, Lemont, IL, 60439, USA.

10 <sup>3</sup> Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

<sup>4</sup> LibraGen, 3 rue des Satellites, 31400 Toulouse, France.

<sup>5</sup> Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.

<sup>6</sup> Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France.

<sup>7</sup>. Centre National de la Recherche Scientifique, UMR8030, 91000 Evry, France.

15 <sup>8</sup>. Université d'Evry Val d'Essonne 91000 Evry, France.

<sup>9</sup> Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.

Running title: Grassland Soil Metagenomics

## 20 Abstract:

The soil ecosystem is critical for human health, affecting aspects of the environment from key agricultural and edaphic parameters to critical influence on climate change. Soil has more unknown biodiversity than any other ecosystem. We have applied diverse DNA extraction methods coupled with high throughput pyrosequencing to explore  $4.88 \times 10^9$  base pairs of metagenomic sequence data from the longest continually studied soil environment (Park Grass experiment at Rothamsted Research in the UK). Results emphasize important DNA extraction biases and unexpectedly low seasonal and vertical soil metagenomic functional class variations. Clustering-based subsystems (CBSS) and carbohydrate metabolism had the largest quantity of annotated reads assigned although less than 50 % of reads were assigned at an E value cutoff of  $10^{-5}$ . In addition, with the more detailed subsystems, cAMP signaling in bacteria ( $3.24 \pm 0.27$  % of the annotated reads) and the Ton and Tol transport systems ( $1.69 \pm 0.11$  % ) were relatively highly represented. The most highly represented genome from the database was that for a *Bradyrhizobium* species. The metagenomic variance created by integrating natural and methodological fluctuations represents a global picture of the Rothamsted soil metagenome that can be used for specific questions and future inter-environmental metagenomic comparisons. However, only 1% of annotated sequences correspond to already sequenced genomes at 96% similarity and E values of less than  $10^{-5}$ , thus, considerable genomic reconstructions efforts still have to be performed.

## Introduction

40 Microorganisms first appeared more than  $3.5 \times 10^9$  years ago (Allwood, 2006, 714-8), ~1.5 billion years after the formation of our planet. Genetic flexibility over a vast expanse of geological time has enabled microorganisms to adapt to virtually every conceivable ecosystem on earth (e.g. Huber et al., 2007; Pointing et al., 2009; Larose et al., 2010). Among contemporary ecosystems, soil, which is a product of microbial and macrobial life, exhibits  
45 the greatest density and phylogenetic diversity per unit volume (Van Elsas et al., 2006; Roesch et al., 2007), with approximately  $10^9$  cells per gram, comprising a diversity that is estimated to range from thousands to millions of taxa (Knietch et al., 2003).

Soil microbial communities are indispensable for the health of our planet; they drive major geochemical cycles (Falkowski et al., 2001) and help to support healthy plant growth (Ortíz-Castro et al., 2009). Yet, there is still a considerable lack of understanding of the mechanisms  
50 of interaction and metabolism that exist among members of the microbial community and their ecosystem. Existing knowledge, concerning the phylogenetic and functional diversity, community metabolic potential, and consequences of evolutionary adaptation, is based largely on partial information gained from studies performed on microorganisms that have been  
55 cultivated from soil on a small scale or 16S rRNA gene sequences.

A dependence on studies of cultivable organisms may limit our fundamental understanding of the diversity of interactions in this system. The organisms cultured from soil so far, represent a fraction of the soil biota, e.g. those amenable to growth in controlled laboratory conditions (Schloss and Handelsman, 2003; Davis et al., 2011). Attempts to apply metagenomic  
60 methodology to soil samples have been hampered by extreme technical challenges, such as extracting an unbiased and representational sample of genetic material from organisms with very different cell membranes and accessible DNA (Delmont et al., 2011b,c; Demaneche et al., 2008; Ginolhac et al., 2004; Handelsman et al., 1998; Rajendhran and Gunasekaran, 2008). This problem is exacerbated by the uneven spatial distribution of microbial  
65 communities in soil (Grundmann, 2004, 119-127; Ranjard, 2001, 707-16}. Unlike marine systems, which are generally well mixed and amenable to temporal and biogeographic observations (Gilbert et al. 2009, 2010), soil systems surveys have, despite a wealth of valuable data acquired from hundreds of well designed experiments and surveys, uncovered only a fraction of the assumed immense microbial diversity of the soil metagenome (e.g.,

70 Tringe et al., 2005; Roesch et al., 2007; Morales et al., 2009). In spite of numerous efforts to study parameters influencing its diversity using cultural independent approaches (e.g., soil pH or nitrogen fertilisation, Rousk et al., 2010; Ramirez et al., 2010), data from soil are scarcer than those collected from other commonly encountered ecosystems. The only contemporary published soil metagenome (Tringe et al., 2005) contains just 100 million base pairs of DNA,  
75 which is potentially a mere millionth of one percent of the genetic material that could be extracted from a gram of soil (based on an assumption of 4 million base pairs per average microbial genome and  $10^9$  cells per gram of soil). The relative lack of available soil related sequence data presents an interesting paradox, that the most diverse environment on earth has received the least attention from metagenomic analysis (Vogel et al., 2009) although the first  
80 soil metagenome dates from 2005. To redress this balance, we have performed an in-depth investigation of a temperate European ungrazed grassland soil metagenome using pyrosequencing technology.

Building on our previous investigations (Delmont et al., 2011b,c), this study describes an unprecedented effort to characterize the microbial diversity and functional potential of a  
85 single soil ecosystem that found in the Park Grass Experiment at Rothamsted Research; the location of the oldest agricultural experiments in the world, run continuously since 1856 (Silvertown et al., 2006). In an attempt to explore this unique environment, almost  $5 \times 10^9$  base pairs of metagenomic sequence data (Titanium pyrosequencing reads) were produced from soils collected from three depths and at three time points spanning two years. To address  
90 concerns regarding the influence of DNA extraction technique bias on microbial diversity (Delmont et al., 2011b), we performed 11 different extraction techniques to improve the diversity of the sequenced microbial genomes. The MG-RAST (Meyer et al., 2008) annotated content of the samples were compared to each other, and to samples of two previously reported, non-soil, environments, so as to place the samples in a more global context.

## 95 **Material and methods**

Soil samples: Samples were collected from the untreated control plot (3d) of Park Grass Experiment, Rothamsted Research, Hertfordshire, UK (Silvertown et al., 2006) in March 2009, July 2009 and July 2010. The overall sample handling is outlined in Figure 1. Soil samples from the top 21 centimeters were collected (Delmont et al., 2011b) by sterile manual  
100 corers (10 cm diameter) in plot 3D at random locations, but not where previous samples had

been taken, and were placed in sterile plastic bags, sealed and placed on ice 24 hours until processing. Previous investigations of this soil demonstrated very little horizontal change in diversity, but measureable changes with depth (Delmont et al., 2011b). Hence, the core samples were fractioned into either seven subsamples as a function of the depth (every three centimeters for the direct lysis (described below) and into two depths for the indirect lysis (Delmont et al., 2011c; described below). The aim of this step is to homogenize the quantity of extracted DNA (which decreases with depth) represented in the final pool for each fraction. The different subsamples were then homogenized separately manually by thorough mixing and stored at -20°C for the direct lysis and at 4°C during a maximum period of one week for the indirect lysis. To access rhizospheric microbial communities, a soil core (0-21cm) was sieved (0.2mm) and grass roots were extracted. Soil attached to roots was then recovered in a water column. The column helped the physical separation between roots and soil present at its surfaces. The few grams of recovered soil were then mixed prior to DNA extraction. The metadata for the site and samples are provided in Table S1.

DNA extraction method: Different extraction procedures were used to process the soil samples (Figure 1). We selected DNA extraction methods that use a wide range of approaches to extract and lyse cells. Among the selected methods, some were already known to provide a high DNA yield (e.g. BIO101), or DNA quality or increased DNA length (in plug lysis), others provided a low yield but could still potentially represent a difficult to access microbial communities. The main goal of this experimental design was to create DNA pools with a large variance in order to uncover a wider range of community members within this soil metagenome at both functional and taxonomical levels.

Direct soil lysis: utilized one of two bead beating protocols, (Fast prep MP Bio101 Biomedical, Eschwege, Germany) (Griffiths et al., 2000) with 0.5 g of soil. This approach was named “direct MP Bio101” (M1, J1, and replicates J1a10 and J1b10). In addition, rhizospheric soil from July 2010 was extracted with the same protocol (J1rhizo10) and soil from July 2009 was extracted with another bead beating method, the MoBio PowerSoil® DNA Isolation Kit (Carlsbad, USA) (J7). Several different indirect DNA extraction methods were used by first extracting cells on a Nycodenz® gradient gel (density of 1.3) (Bertrand et al., 2005) and then applying one of the following four lyses with the extracted cells: 1) the same bead beating protocol, called “indirect MP Bio101” (replicates M2a and M2b from March 2009); 2) the Nucleospin® Tissue kit, named “indirect DNA Tissue” (M4 and J4,

March and July 2009, respectively); 3) the Gram positive kit, named “indirect Gram positive” (M5); and finally 4) a lysis using agarose plugs called “indirect lysis in plug”. (M3 and M6  
135 from 0 to 10cm and 11 to 21 cm depths, March 2009, respectively – see figure 1). Plugs were first transferred in 3 ml of G<sup>-</sup> lysis buffer (1% lauroyl sarcosine, 500 mM of EDTA Na<sub>2</sub>, pH 9.5) with 0.5 mg/ml of lysozyme and incubated at 37°C for 12 h. The agarose plugs were then incubated in 3 ml of G<sup>-</sup> lysis buffer with 500 µg/ml of proteinase K at 56°C for 12 h, and finally equilibrated in a 10 mM Tris (pH 8.0), 1 mM EDTA storage buffer). This enzymatic  
140 lysis was performed in a stable environment (the agarose plug) and was performed without any physical perturbations (*e.g.*, tube mixing that break DNA). This method is generally used to provide high quality and long DNA sequences for the construction of fosmid libraries or for genome size. General information about the different DNA extraction yields used is presented in the table 1.

145 Pyrosequencing runs: A minimum of 10 µg of DNA were used for each Roche/454 pyrosequencing run on a 454 pyrosequencer (GS FLX Titanium Series Reagents ; Roche 454; Shirley, NY, USA). Processing of samples (prior to sequencing) did not involve prior amplification step. For the direct lysis, equal quantities of DNA extracted from the seven fractions from 0 to 21cm were pooled together. J1a10 and J1b10 correspond to distinct  
150 extractions from the same soil core. For the indirect approach corresponding to soil from 0 to 21cm, equal quantities of DNA extracted from the two fractions (0 to 10 and 11 to 21cm) were pooled together. For the indirect lysis using the bead beating protocol (0 to 21cm, March 2009), two pyrosequencing runs (M2a and M2b) were performed from the same DNA pool (> 20 micrograms). The sequence data are publically available  
155 (<http://www.genomenviron.org/Projects/METASOIL.html>).

Data analyses: Artificial duplicates were deleted using cd-hit-454 with default parameters (Niu et al. 2010). Sequences were then annotated on the MG RAST (v.02) online software (Meyer et al., 2008). Reads were distributed into different metabolic subsystems. Similarity search between pyrosequencing reads and the SEED database (Overbeek et al., 2005) have  
160 been processed with a maximum E value of 10<sup>-5</sup>. All compared distributions were normalized as a function of the number of annotated sequences for each metagenome. Data corresponding to both functional and taxonomical distributions were then statistically analyzed within the STAMP software (Parks and Beiko, 2010). Fisher’s exact tests were performed and annotated

functions and taxa with p-values < 0.05 were considered to be significantly different between  
165 the different experiments

Tests on assembly productivity were performed using Newbler (Margulies et al., 2005).  
Newbler was run directly from the “.sff” files produced by the pyrosequencer using the  
following parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed  
step: 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap  
170 identity : 90%; Alignment identity score: 2; Alignment difference score: -3. The minimum  
read length in the data set was 40. Each deeply sequenced dataset was assembled separately  
using the 454 GS de novo assembler software (Newbler v2.0.00.22.), and all contigs were  
used for subsequent analysis. In addition, MetaGeneMark (version 2.7d using the parameter  
file for metagenome gene prediction version 1) have been used to search genes from the 100  
175 largest contigs, and the 1006 genes predicted were analyzed via MG-RAST.

## Results

Thirteen pyrosequencing runs were performed with DNA extracted from the Rothamsted  
Research (Park Grass) site. Grassland soil samples were taken at different depths and three  
180 different time points over 1.5 years. DNA was extracted from the samples using 6 different  
DNA extraction protocols (see materials and methods and figure 1). Two samples were  
sequenced in duplicate (J1a10 and J1b10, and M2a and M2b) to explore the reproducibility of  
the metagenomic profile. A total of 12,575,129 reads were generated (length average of 385.9  
 $\pm$  31.8 bp) and 34.5 ( $\pm$ 3.3) % of them were annotated with the MG-RAST online server (E  
185 value <  $10^{-5}$ ) (Meyer et al., 2008). Based on the protein database used by MG-RAST, 88.64  
( $\pm$ 1.44) % of these annotated sequences had closest homology to a protein found in Bacteria,  
0.91 ( $\pm$ 0.23) % to Eukarya, and 1.41 ( $\pm$ 0.16) % to Archaea. Thus, almost 9% of annotated  
sequences were not classified at the domain level. All the annotated reads were compared to  
SEED-NR, FIGFams for functional assignments and then used in subsystem reconstructions.  
190 The closest matched gene was the source of information about the functional (metabolic)  
subsystem that the read was binned into and about the “taxa” represented by this read.  
Therefore, the taxa cited here correspond to the genomes in the database that best matched the  
given read as long as the E value was smaller than  $10^{-5}$ . Major functions and taxa identified  
can be found in tables S2 through S4.

195 Functional comparison:

Functional differences between the 13 datasets generated from Rothamsted, two datasets from other soils, and one from an aquatic environment were derived by exploring the relative number of reads associated with the 835 functional subsystems detected at least in one metagenome (Figure 2). Bootstrap values are provided. The method of DNA extraction correlates with sample grouping. Samples, M1, J1, J1.a and J1.b were directly extracted using the MP Bio101 kit. Sample J7 which lies in the same general group was extracted directly with MoBio Powersoil kit. The sample from the application of direct MP Bio101 on the rhizosphere soil (J2) is closely associated with this group. The bootstrap values are not particularly high within this group. Three sample pairs on the other hand had significant bootstrap values (>90%) grouping them apart from the other samples: 1) the replicate samples from the application of MP Bio101 to the cells first removed from soil via the Nycodenz gradient (M2.a and M2.b); 2) the two depth samples extracted by indirect lysis in agarose plugs (M3 and M6); and 3) the two samples from different seasons extracted after Nycodenz by use of the DNA tissue kit (M4 and J4). In order to assess the statistical likelihood of the subsystem distribution differences between samples, STAMP software (Parks and Beiko, 2010) based on a bootstrap approach using Fisher's exact tests were applied to the MG-RAST (subsystem functional level 3) outputs. This approach determined what percentage of the 835 subsystems were significantly (at 95% CI) different between any pair-wise comparison. Replicate runs (M2a / M2b and J1a10 / J1b10) had between 7.3 and 7.7 % dissimilar subsystems and seasonal variations had 8.6 and 11.7 % dissimilar subsystems for direct (M1/J1) and indirect (M4/J4) extractions respectively. When different lysis methods (e.g., M4, M5, and M6) were applied to the bacterial cells removed by Nycodenz gradient gel before DNA extraction, significant differences in subsystem distributions (16.9 – 39.8% dissimilar subsystems) were observed at the 95% CI. Using sequences corresponding to communities extracted from two distinct horizons (0 to 10cm: M3 and 11 to 20cm: M6), 27.01% of the detected functional subsystems possessed statistically different distributions. Two types of geographical comparisons were made. One was between Rothamsted soil and soil from Italy (Vallombrosa forest soil, defined as a Cambic Umbrisol) extracted with the same method (MP Bio101) in our laboratory (the related Italian soil metagenome is represented by approximately 100 000 sequences) and these two soils had 14.1% dissimilar subsystems. The second was Rothamsted sequences compared to those from Puerto Rico

(located in the Luquillo experimental forest and defined as a tropical rain forest soil, Metagenome ID of 4446153.3 on MG RAST, one million reads), which were extracted and sequenced elsewhere. They had between 30.98% and 33.13% dissimilar subsystems. The most extreme comparison was between Rothamsted soil and the Sargasso Sea (72% dissimilar subsystems) as indicated also by the distance in figure 2.

Among the major (29) metabolic classes, clustering-based subsystems (CBSS) and carbohydrate metabolism had the largest quantity of annotated reads assigned (Figure 3). Virulence and amino acid and derivatives were next in prevalence (Figure 3). The cluster-based subsystems contain such functions as proteosomes, ribosomes and recombination-related clusters. The virulence subsystem contains diverse functions also, such as resistance to antibiotics and toxic compounds, and pathogenicity islands. Some subsystems were relatively minor such as photosynthesis, prophage, dormancy and sporulation (Figure 3). Although there was significant (at the 95%CI) differences in the distribution of reads in some (from about 7 to 40%) of the different metabolic subsystems from the different pyrosequencing runs of DNA extracted from Rothamsted soil, the standard deviation around the mean of all of the pyrosequencing runs varied between 2 and 50 %, with the higher variance for the metabolic classes with relative few assigned reads (*e.g.*, macromolecular synthesis; error bars in figure 3). When comparing the assigned reads at a finer functional subsystem classification within MG-RAST, the most prevalent subsystem (out of the 835 different categories) in the soil was the cAMP signaling in bacteria with  $3.24 \pm 0.27$  % of the annotated reads (Table S2). The next most prevalent subsystem was the Ton and Tol transport systems at  $1.69 \pm 0.11$ % of the annotated reads (Table S2). These prevalent systems varied less than 10% between DNA extraction pools except for the distribution of CO<sub>2</sub> uptake carboxysome related genes, which varied from 0.56% in M3 to 1.43% in M4, which represents an increase of 60.7% (average for the thirteen pyrosequencing runs was  $0.70 \pm 0.42$  % of reads).

### Taxonomic comparison:

The  $56 \pm 4.4\%$  of annotated protein sequences showed closest homology to a total of 1214 unique taxa using the taxonomic annotation of functional SEED subsystems. The most dominant putative taxon was *Solibacter usitatus* ( $6.72 \pm 0.29\%$  of annotated reads). Other taxa with relatively high number of assigned reads were *Blastopirellula marina* ( $4.96 \pm 2.88\%$ ), *Bradyrhizobium japonicum* ( $4.89 \pm 0.635\%$ ) and *Acidobacteria bacterium* ( $3.64 \pm 0.94\%$ ) (Table S3). The legitimacy of the read assignment at an E value of  $10^{-5}$  cut-off is provided in part by the E value distribution of the different reads assigned to the reference genome. In the case of *Bradyrhizobium japonicum*, the majority of assigned reads had E values lower than  $10^{-30}$  and in the case of *Blastopirellula marina*, the E values were in general larger than  $10^{-30}$ . Using the taxonomic classification of functional gene fragments, it is possible to use all annotated reads to determine community structure (Figure 4); however, it is also possible to use 16S rRNA sequences to determine the community structure, although the number of reads is considerably less than for the SEED annotation. Three different databases accessible within the MG-RAST platform and used with the MG-RAST software were used to determine community structure with standard deviations calculated from the variance of the 13 different pyrosequencing runs (Figure 4). While there is a general agreement: alpha-, beta- and gammaproteobacteria and Actinobacteria dominate all four methods, there are some important differences in the relative number of reads in different classifications. For example, the Silva SSU database 94 has a much higher percentage of reads in Flavobacteria than the other systems (Figure 4). In order to use functional genes other than 16S rRNA for taxa or at least genera identification, a more accurate and limited analysis constrained the similarity at 96% or better (still with an E value of  $10^{-5}$ ). When this was performed using SEED, only  $0.35 \pm 0.09\%$  of the total reads (or about 1% of the annotated reads) were used to identify bacterial taxa (Table S4). The most abundant taxa identified from Rothamsted soil were members of the *Bradyrhizobium*, *Rhodopseudomonas* and *Nitrobacter* genera (Alphaproteobacteria); the *Solibacter* and *Acidobacteria* genera (Acidobacteria) and *Pseudomonas* (Gammaproteobacteria) and *Burkholderia* (Betaproteobacteria) genera. *Blastopirellula marina* was no longer associated with any of the reads.

### Soil metagenome assembly:

Sequence data were assembled to provide a metric describing the depth of sequencing applied to the community metagenome; in part this was used to estimate the minimum quantity of sequencing required to completely sequence all the members of the soil microbial community. The extreme minimum could be considered as the quantity of sequences where no singleton is left unassembled, even if practically this minimum is insufficient to assemble all the genomes. Ten random read subsamples of increasing metagenome size (read quantity) were run through the Newbler assembler (Table S5). No attempt was made here to optimize the assembly process. The ten subsamples ranged in size from 1257242 to 12572342 reads (with 487554794 to 4874169257 total number of bases) and produced from 7478 to 266600 contigs (Table S5). The largest contig size increased from 6361 bp to 22645 bp with three times as many reads but then decreased and leveled off at about 15400 bp with increasing number of reads (Table S5). The fraction of reads that were not included in any contig (“singletons”) fell from roughly 0.93 to 0.76 when increasing the number of reads ten-fold (Figure 5A insert). This data was fitted and extrapolated to the point where no read would be orphaned. This extrapolation was at about 400 million 454 reads (average of 386 bp in length) with the 95% confidence interval stretching from less than 200 million reads to almost 1400 million reads (Figure 5A). The maximum contig length did not continue to increase with increasing read number, but the number of reads per contig did develop two general trends (Figure 5B). These two trends are schematically represented by the two lines in figure 5B. The denser trend has a slope represented by a contig coverage of about 30x (when the assembler needs/uses 30X to build the contigs) and the smaller trend has a contig coverage of about 4.5x. Contigs from these two trends were selected, broken in coding sequences by MetaGeneMark and then annotated using MG RAST. Globally, the trend corresponding to coverage of 30x possessed more sequences related to Firmicutes (10.99%) and Verrucomicrobia (21.85%). In contrast, the trend corresponding to low coverage assembled contigs (4.5x coverage) possessed a majority of sequences related to Proteobacteria (66.06%). Independent of the two observed trends, the 100 largest contigs created from the entire sequence pool were also annotated by MG RAST and in general the relative proportion of different functional and phylogenic classes (stars in figures 3 and 4) were similar to that for the sequences directly with some exceptions. There were fewer virulence subsystem hits and significantly more fatty acids and protein metabolism hits.

## Discussion

Soil is one of the most diverse environments on earth and the depth of the microbial diversity  
315 is still poorly understood. High throughput sequencing technologies, coupled with appropriate  
DNA extraction methods, provide a means to explore the soil ecosystem with an  
unprecedented level of detail (Vogel et al, 2009). In this study, pyrosequencing from 13  
samples generated nearly  $5 \times 10^9$  base pairs of sequence data with average read size of 386 bp.  
Three key parameters were varied: soil depth, sample collection season, and DNA extraction  
320 method. Sequence samples were annotated with the MG-RAST online server, revealing broad  
functional (835 of 878 possible functional subsystems) and taxonomic (detection of 1214  
putative taxa) diversity in the Rothamsted Park Grass soil metagenome.

The most abundant functional subsystems in the Rothamsted soil were seemed to be related to  
microbial cAMP signaling and Ton and Tol transport (Table S2). The same subsystems were  
325 prevalent in metagenomes in soil at Waseca farm, in Puerto Rico and Italy. These trends in  
soil functional content are robust enough to be observed on a global scale. cAMP is an  
important secondary messenger in Eukarya and Bacteria. cAMP is a universal cell  
energy/metabolism regulator as well as being involved with cell-cell signaling. Soil bacteria  
might have to deal with frequently fluctuating substrate levels so that they would need extra  
330 regulation rather than interacting with plants. Interestingly, since cAMP is also a subversion  
mechanism, some bacterial pathogens might also subvert plant cAMP production for their  
own benefit, through injection of adenylate cyclase and/or various toxins that alter adenylate  
cyclase levels (adenylate cyclase is essential to the production of cAMP) (Agarwal et al., 2009;  
Akhter et al., 2008). Iron is an essential element for most organisms (Weinberg, 1984), but  
335 can be a limiting reagent for life (often in oceans, Boyd et al, 2007) due to its insolubility in  
aerobic environments at neutral pH. In response to this stress, some bacteria possess high-  
affinity transport systems (Crosa et al., 2004) and generate high-affinity siderophores that  
complex extracellular iron (Neilands et al, 1980) to optimize its acquisition. The presence of  
Ton related proteins in the soil is likely due to TonB, an energy dependent cell envelope  
340 protein that assists iron uptake through accommodation of ferric siderophores, too large to  
cross porins, through the outer membrane (Klebba et al, 2003).

MG-RAST annotation also revealed the presence of several highly abundant cluster-based  
subsystems (CBSS). These are groups of functionally coupled genes (genes found proximal to  
each other in the genomes of diverse taxa) whose functional attributes are not well  
345 understood. The relatively high abundance of these subsystems across all Park Grass samples,

as well as the other sequenced soils, suggests that they play key roles in soil ecosystems across the globe, and should be explored in future research efforts to understand the composition of soil ecosystems. The CBSS-258594.1.peg.3339, CBSS-269799.3.peg.2220, CBSS-83332.1.peg.3803, CBSS-249196.1.peg.364 (Table S2) are thought to be a  
350 galactoglycan biosynthesis, a molybdenum oxidoreductase, a PKS-related, and a fatty acid metabolism subsystem, respectively.

The comparison of the runs corresponding to the same DNA sample (M2a/M2b) provided important information about the reproducibility of pyrosequence generation in highly biodiverse environments. The Fisher's exact test operated by the STAMP software did  
355 identify some functions (about 7%) and taxa that varied significantly (at the 95% CI) between replicates. The lower p-value was on the order of  $10^{-7}$  when comparing M2a and M2b at the functional level, so some comparisons between seasons and depths were possible. Based on these observations, functional comparisons having at most a minimum p-value of  $10^{-8}$  (cut-off based on the observed technological reproducibility) were considered to have distributions  
360 that varied significantly. Unfortunately, the technological reproducibility is not the only limit for robust metagenomic comparisons. Even if a stringent p-value is used, the DNA extraction approach influenced the experimental conclusions. When comparing the seasonal effect by using two different extraction approaches (direct:M1/J1 and indirect M4/J4), some differences in relative predominance of different subsystems were found. Based on the comparison of M1  
365 and J1, sequences related to the type 4 secretion and conjugative transfer and cellulosome subsystems are more represented in March (p-value of  $10^{-8}$  in the two cases). When comparing M4 and J4, the cellulosome subsystem is still detected more in March (p-value  $<10^{-15}$ ), but the type 4 secretion and conjugative transfer is not. In contrast, sequences related to bacterial cAMP signaling are more present in July (p-value of  $10^{-12}$ ), but only when comparing M4 and  
370 J4. Thus, only sequences related to cellulosome dominated one season's metagenome independent of the extraction method applied. Major environmental difference between the two studied seasons was temperature (from 6°C in March to 16.6°C in July). In addition, snow lay on the ground for weeks in February of the same year, thus limiting active grass growth. As a consequence, soluble root exudates were possibly in short supply during this relatively  
375 cold period and cellulosome from root residues would be the main source of carbon and energy supporting soil microbial communities.

On the other hand, depth had more effect with sequences related to genes involved in bacterial chemotaxis, Ton and Tol transport systems, flagellum mechanism, D-ribose and L-Arabinose utilization represented more in the surface sample (0 to 10 cm) and sequences related to selenocysteine metabolism and tRNA aminoacylation represented more at depth (11 to 20 cm). However these results were generated using only one DNA extraction method. In comparison to depth and seasonal variables, the extraction method was able to influence functional distributions (Figure 2), especially when using methods with striking differences in cell lysis (e.g., Gram positive kit versus in agarose plug lysis or DNA tissue). Thus, the stringency of lysis appears to be a crucial step for soil metagenomic analysis, confirming previous results with RISA and phylogenetic microarray analyses (Delmont et al, 2011b).

In addition, when studying the distribution of sequences based on their G+C%, clear variations were found among the different runs. Direct lysis versus indirect lysis had more impact on the G+C% profile than any other variable. The indirect lysis provided more sequences possessing a higher G+C ratio (from 60 to 72%), while the direct lysis had a more even distribution with more sequences in the 50 to 58 G+C% range (Figure S1). Both metagenomic standard deviations and G+C% ratio profile fluctuations are limited by the experiments and variables used. However, this effort provides both significant soil metagenomic sequences and data useful to appreciate methodological differences in microbial community diversity accessibility.

Given the relatively low functional subsystem variations between different soils (figure 2), soil microbial community metagenomes from Rothamsted, Puerto Rico, Italy and the Waseca farm soil (Tringe et al., 2005) could be compared to metagenomes from oceans and human feces. This comparison might help identify some of the soil ecosystem unique functional attributes. In order to make the comparison, principal component analysis was generated based on the distribution of general functional subsystem classes with metagenomes publically available from these ecosystems (Figure 6). Some general functional classifications appear to be relatively more represented in one ecosystem in comparison to the others. Sequences related to RNA and protein metabolism, photosynthesis, fatty acids and lipids, and macromolecular synthesis are more highly represented in ocean metagenomes. In contrast, phosphorus metabolism and virulence are less represented in ocean metagenomes than in those sequenced for soil and human microbiomes. Sulfur and potassium metabolism, membrane transport, stress response and regulation, and cell signaling are more represented,

and nucleosides and nucleotides, and RNA and protein metabolism are less represented in soil  
410 metagenomes. In human microbiomes, cell division and cell cycle, DNA and phosphorus  
metabolism, cell wall and capsule, dormancy and sporulation, carbohydrates are more  
represented than in those of oceans and soils (Figure 6). When comparing the taxonomical  
structure of these metagenomes, Cyanobacteria and Bacteroidetes appear to be more  
415 represented in the oceans. In addition, Eukaryotic sequences were also detected and represent  
additional specificities of these metagenomes (Figure S2). Actinobacteria, Chloroflexi,  
Fibrobacteres and Acidobacteria group, Planctomycetes, and Synergistetes are more present in  
soils. Chlorobi, Firmicutes, Spirochaetes, Fusobacteria and the Bacteroidetes Chlorobi group  
are clearly relatively dominant in human digestive tracts. In contrast, Proteobacteria are more  
present in oceans and soils. The metagenomes are clearly grouped as a function of the  
420 environment based on both general functional and taxonomical distributions. So in spite of  
important DNA extraction biases and sequencing technology differences (Illumina,  
Pyrosequencing and Sanger), global metagenomic comparisons are possible and provide  
unique information about the functional and taxonomical differences of each environment  
(Delmont et al., 2011a). As an example, sequences related to metabolism of aromatic  
425 compounds are more abundant in soils possibly due to the presence of these compounds in  
this environment. However, additional comparisons, such as qPCR and metatranscriptomics,  
need to be performed to confirm which taxa and functions are unusually active in soil to gain  
a better understanding of soil microbial community function.

The relative percentage of orphan reads decreased continually when accumulating  
430 pyrosequences. Therefore, an estimate of the number of reads needed to avoid having orphan  
reads would possibly provide the absolute minimum number of reads needed to sequence the  
entire soil metagenome. Rarefaction analysis of this sequencing effort (Figure 5) indicated  
that the equivalent of about 320 Titanium runs would be required to create contigs from all of  
the soil pyrosequence reads generated. Of course, chimeras might be generated due to the  
435 complexity of communities, and a much larger effort would be needed to assemble the soil  
metagenome, but as new efficient high-throughput sequencing technologies and valuable  
assembling tools are developed, this goal will become less utopic. Genomes from  
Proteobacteria might be assembled more rapidly than those from Firmicute or  
Verrucomicrobia phyla. The presence of regions that limit assembly (e.g., insertion sequences  
440 regions) and the complexity of diversity among taxa might explain in part the efficiency

differences observed between these phyla (4.5x and 30x), but additional experiments are needed to understand the two trends observed in the figure 5B.

#### Conclusion:

In this study, more than 12 million reads were generated from the soil of the Rothamsted  
445 Research Park Grass experiment. These sequences were generated in 13 separate sequencing  
runs producing over  $4 \times 10^9$  bp. The results demonstrated both some DNA extraction biases  
and relatively low seasonal (when comparing March and July months) and vertical soil  
metagenomic functional class fluctuations. In addition, this approach provided a statistical  
450 view of functional distributions in this soil. This metagenomic study increased our knowledge  
about soil microbial communities at a metagenomic level by integrating both natural and  
methodological fluctuations. The metagenomic variance so generated represents a global  
picture of the Rothamsted soil metagenome that can be used for specific questions and future  
inter-environmental metagenomic comparisons. However, only 34.5 % of the reads were  
455 assigned to functions and less than 1% of annotated sequences correspond to already  
sequenced genomes (at 96% similarity), therefore, many soil microorganisms remain elusive  
and genome constructions are needed.

Acknowledgements: T.O.D. was supported by the Rhône-Alpes Region. We want to thank  
the French National Research Agency (ANR) for financing Metasoil (Projet ANR-08-GENM-  
025) and the European Union (7<sup>th</sup> Framework KBBE-2007-3-3-05) funding for Metaexplore  
460 (22625) project. This work was supported in part by the U.S. Dept. of Energy under Contract DE-  
AC02-06CH11357.

#### References

- Agarwal N and Bishai WR. (2009). cAMP signaling in Mycobacterium tuberculosis. *Indian J  
Exp Biol* 47:393-400.
- 465 Akhter Y, Yellaboina S, Farhana A, Ranjan A, Ahmed N, Hasnain SE. (2008). Genome scale  
portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in  
pathogenesis. *Gene* 407:148-58.
- Allwood AC, Walter MR, Kamber BS, Marshall CP, Burch IW. (2006). Stromatolite reef  
from the Early Archaean era of Australia. *Nature* 441:714-718.

- 470 Davis KE, Sangwan P, Janssen PH. (2011). Acidobacteria, Rubrobacteridae and Chloroflexi are abundant among very slow-growing and mini-colony-forming soil bacteria. *Environ Microbiol* 13:798-805.
- Delmont TO, Malandain C, Prestat E, Larose C, Monier J-, Simonet P, et al. (2011a). Metagenomic mining for microbiologists. *ISME Journal*. DOI: 10.1038/ismej.2011.61
- 475 Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P et al. (2011b). Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* 77:1315-24.
- 480 Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. (2011c). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods*. 2011;86(3):397-400.
- Demaneche S, Sanguin H, Poté J, Navarro E, Bernillon D, Mavingui P et al. (2008). Antibiotic-resistant soil bacteria in transgenic plant fields. *Proc Natl Acad Sci USA* 105: 3957-3962.
- 485 Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM et al. (2008). Functional metagenomic comparison profiling of nine biomes. *Nature*. 452:629-632.
- Falkowski PG. (2001). Biogeochemical cycles. *Encyclopedia Biodivers*. 1:437-453.
- Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* 105:3805-10.
- 490 Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tophile K et al. (2004). Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl. Environ. Microbiol* 70:5522-5527.
- 495 Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998). Molecular Biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 5:245-249.

- Huber JA, Pointing SB, Chan Y, Lacap DC, Lau MC, Jurgens JA, Farrell RL. (2009). Highly specialized microbial diversity in hyper-arid polar desert. *Proc Natl Acad Sci USA* 106:19964-9.
- 500 Kahvejian A, Quackenbush J, Thompson JF. (2008). What would you do if you could sequence everything? *Nat Biotechnol* 26:1125-1133.
- Knietch A, Waschowitz T, Bowien S, Henne A, Daniel R. (2003). Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on *Escherichia coli*. *J Microbiol Biotechnol* 5: 505 46-56.
- Larose C, Berger S, Ferrari C, Navarro E, Dommergue A, Schneider D, Vogel TM. (2010). Microbial sequences retrieved from environmental samples from seasonal Arctic snow and meltwater from Svalbard, Norway. *Extremophiles* 14:205-12.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA *et al.* (2005). Genome 510 sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-80.
- Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The Metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes *BMC Bioinformatics* 9:386.
- Morales SE, Holben WE. (2009). Empirical testing of 16S rRNA gene PCR primer pairs 515 reveals variance in target specificity and efficacy not suggested by *in silico* analysis. *Appl Environ Microbiol* 75:2677-83.
- Nealson KH, Venter JC. (2007). Metagenomics and the global ocean survey: what's in it for us, and why should we care? *ISME J* 1:185-7.
- Niu B, Fu L, Sun S, Li W. (2010) Artificial and natural duplicates in pyrosequencing reads of 520 metagenomic data. *BMC Bioinformatics* 11:187.
- Ortiz-Castro R, Contreras-Cornejo HA, Macías-Rodríguez L, López-Bucio J. (2009). The role of microbial signals in plant growth and development. *Plant Signal Behav* 4:701-12.

- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000  
525 genomes. *Nucleic Acids Research*, 33:5691-5702.
- Parks DH, Beiko RG. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26:715-21.
- Rajendhran J, Gunasekaran P. (2008). Strategies for accessing soil metagenome for desired applications. *Biotech Adv* 26: 576-90.
- 530 Ranjard L, Richaume AS. (2001). Quantitative and qualitative microscale distribution of bacteria in soil. *Res Microbiol* 152: 707–716.
- Ramirez KS, Lauber CL, Knight R, Bradford MA, Fierer N. (2010). Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. *Ecology*91:3463-70.
- 535 Roesch LL, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD et al. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* 1: 283-290.
- Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J* 4:1340-51.
- 540 Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464:59-65.
- Schloss PD, Handelsman J. (2003). Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* 14: 303-310.
- Shendure J, Ji H. (2008). Next-generation DNA sequencing. *Nat Biotech* 26: 1135-1145.
- 545 Silvertown, J, Poulton P, Johnston E, Edwards G, Heard M, Biss PM. (2006) The Park Grass Experiment 1856-2006: its contribution to ecology. (2006). *Journal of Ecology* 94: 801–814.
- Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW et al. (2005). Comparative Metagenomics of Microbial Communities. *Science* 308:554-557.
- Van Elsas JD, Jansson JK, Trevors JT. (2006). *Modern Soil Microbiology II*, CRC press.

550 Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD et al. (2009).  
TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Microbiol*  
7:252.

Willner D, Thurber RV, Rohwer F. (2009). Metagenomic signatures of 86 microbial and viral  
metagenomes. *Environ Microbiol* 11:1752-66.

555

555 **Figure and Table Legends**

**Table 1.** Quality and quantity of DNA extracted from the Rothamsted Parkgrass soil with different DNA extraction approaches.

**Figure 1.** The sampling and DNA extraction schematic for the thirteen pyrosequencing runs. The two pairs, M2a/M2b and J1a10/J1b10, are respectively replicate runs from the same DNA  
560 extraction and distinct DNA samples extracted sequentially from the same soil sample.

**Figure 2.** Cluster tree confronting the thirteen pyrosequencing runs, two other soil metagenomes and a metagenome corresponding to Sargasso Sea environment based on the number of reads assigned to each of the 835 metabolic subsystems detected by MG-RAST at least in one dataset. The tree was constructed using Euclidean distances, nPCA ordination  
565 method, and complete cluster method.

**Figure 3.** Relative distribution (in percentage of annotated reads) of the 29 major metabolic subsystems (using SEED subsystems in the MG-RAST program) detected in the Rothamsted soil metagenome. Standard deviations correspond to the variability among sequencing runs. The stars represent the relative distribution among the 100 largest contigs after assembly.

**Figure 4.** Relative distribution of microbial classes in the Rothamsted soil metagenome. Standard deviations correspond to the fluctuation of the relative distribution between different pyrosequencing runs. The total number of reads annotated by the different methods is not the same as the SEED annotation using all annotated reads and the others use only identified 16S rRNA genes (*rrs*). The version of Greengenes database used within MG-RAST was from  
570 2008. The stars represent the relative distribution among the 100 largest contigs after assembly based on SEED annotation.

**Figure 5.** Panel A. Relation between number of 454 sequence reads used in the Newbler assembler and the percentage of reads not combined with any other reads (singletons). A best fit equation for this relationship is:  $p\text{Singleton} = a * [\text{nbReads}]^b + c$  with the following four  
580 parameters: Estimated value, Std. Error, t value,  $\text{Pr}(>|t|)$  - for a:  $-6.714 \times 10^{-4}$ ,  $5.409 \times 10^{-5}$ ,  $-12.41$ ,  $5.06 \times 10^{-6}$ ; for b:  $3.703 \times 10^{-1}$ ,  $4.446 \times 10^{-3}$ ,  $83.30$ ,  $9.46 \times 10^{-12}$ ; for c:  $1.047$ ,  $2.372 \times 10^{-3}$ ,  $441.56$ ,  $< 2 \times 10^{-16}$ .

Panel B. Plot of the number of reads per contig as a function of the length of the contigs produced with all the reads from the 13 pyrosequencing runs using the 13 pools of DNA extracted from the Park Grass soil at Rothamsted Research.

585

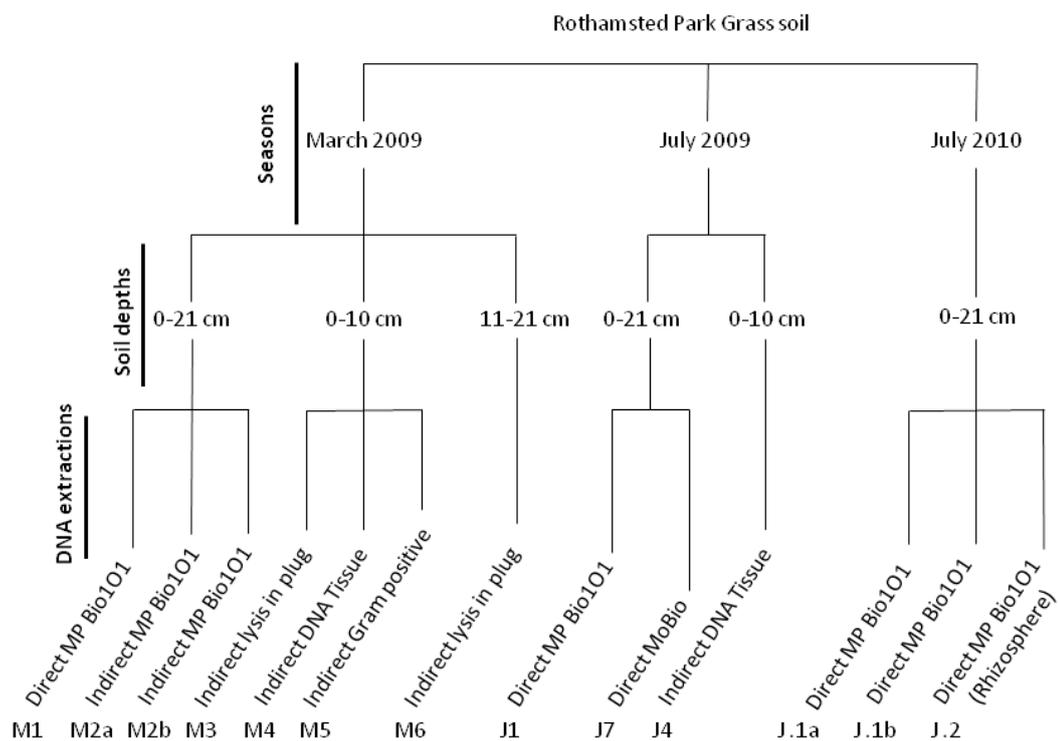
**Figure 6.** The principal component analysis of three ecosystems using the relative distribution of reads in the different metabolic subsystems for the metagenomic sequences available in the public database in addition to those produced here. The large metabolic classes as determined by MG-RAST are mapped on the same PCA as the ecosystems.

590

590 Table 1. Quality and quantity of DNA extracted from the Rothamsted Parkgrass soil with different DNA extraction approaches.

	Quantity of soil used	Principal type of lysis	average DNA length after extraction	DNA yield per kilogram of soil
<b>MP BIO1O1 rhizosphere soil</b>	0.5 g	Mechanical	10 kbp	40 mg
<b>MP BIO1O1 soil</b>	0.5 g	Mechanical	10 kbp	10 mg
<b>MoBIO soil</b>	0.5 g	Mechanical	10 kbp	2 mg
<b>MP BIO1O1 on extracted cells (Nycodenz)</b>	300-400 g	Mechanical	10 kbp	150 µg
<b>In plug on extracted cells (Nycodenz)</b>	300-400 g	chemio-enzymatic	> 500 kbp	120 µg
<b>DNA Tissue on extracted cells (Nycodenz)</b>	300-400 g	chemio-enzymatic	20-40 kbp	30 µg
<b>Gram positive on extracted cells (Nycodenz)</b>	300-400 g	chemio-enzymatic	20-40 kbp	5 µg

Figure 1



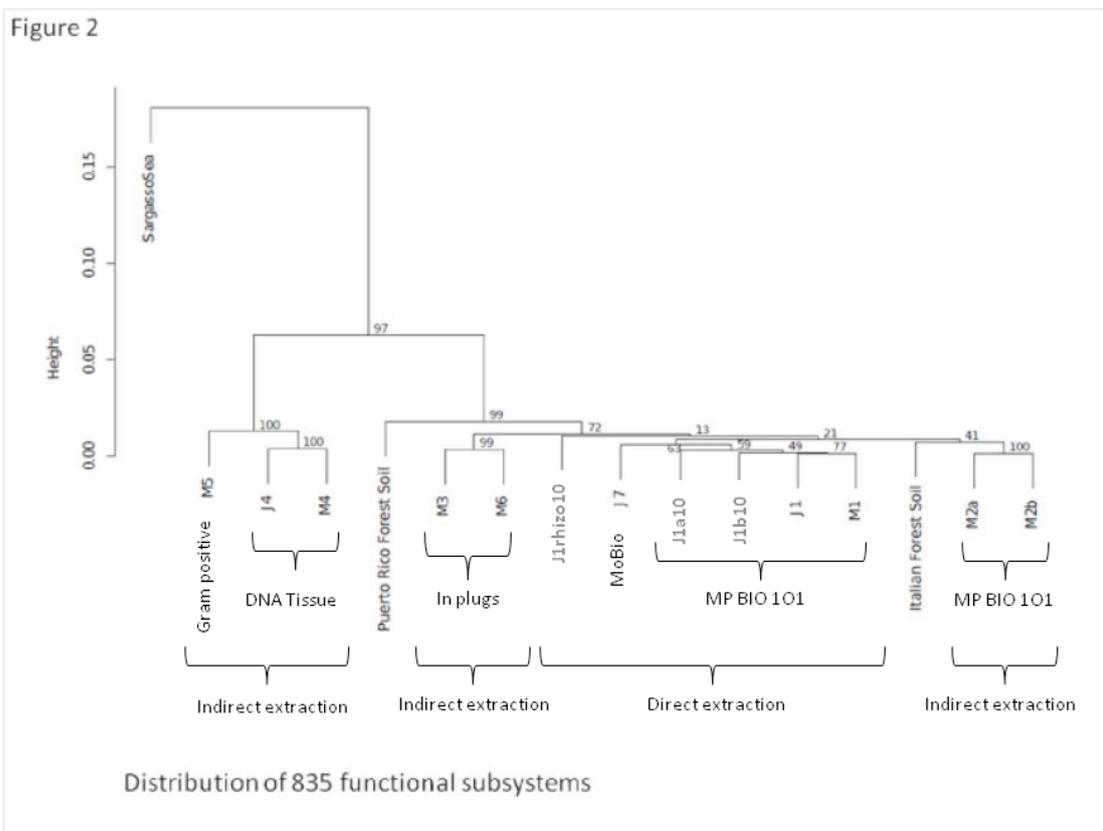


Figure 3

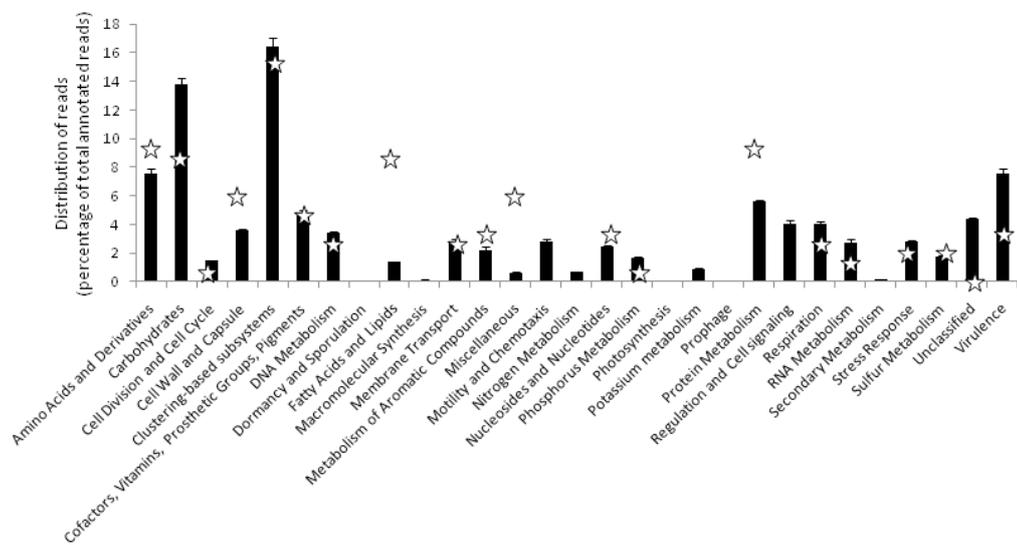


Figure 4

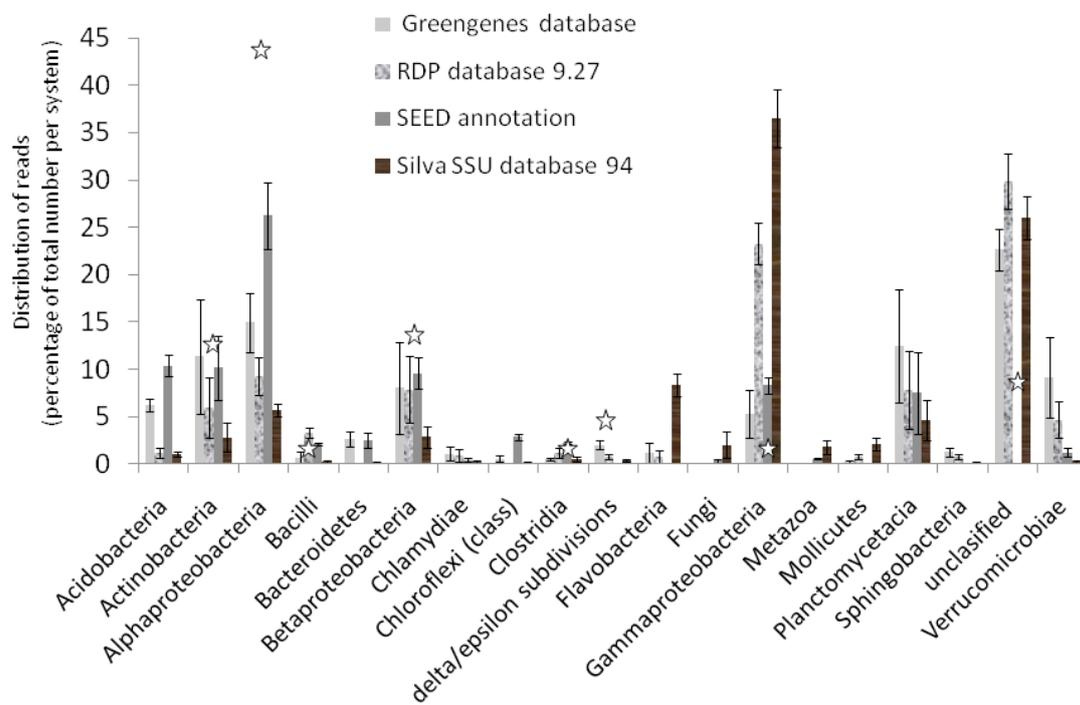
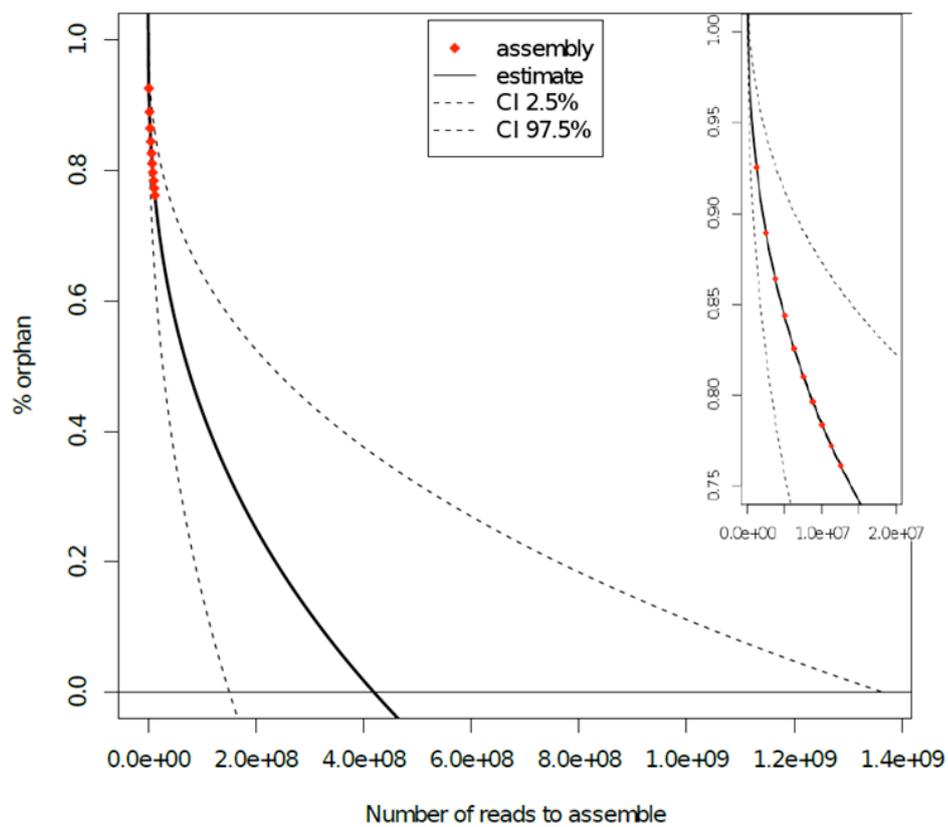


Figure 5  
Panel A

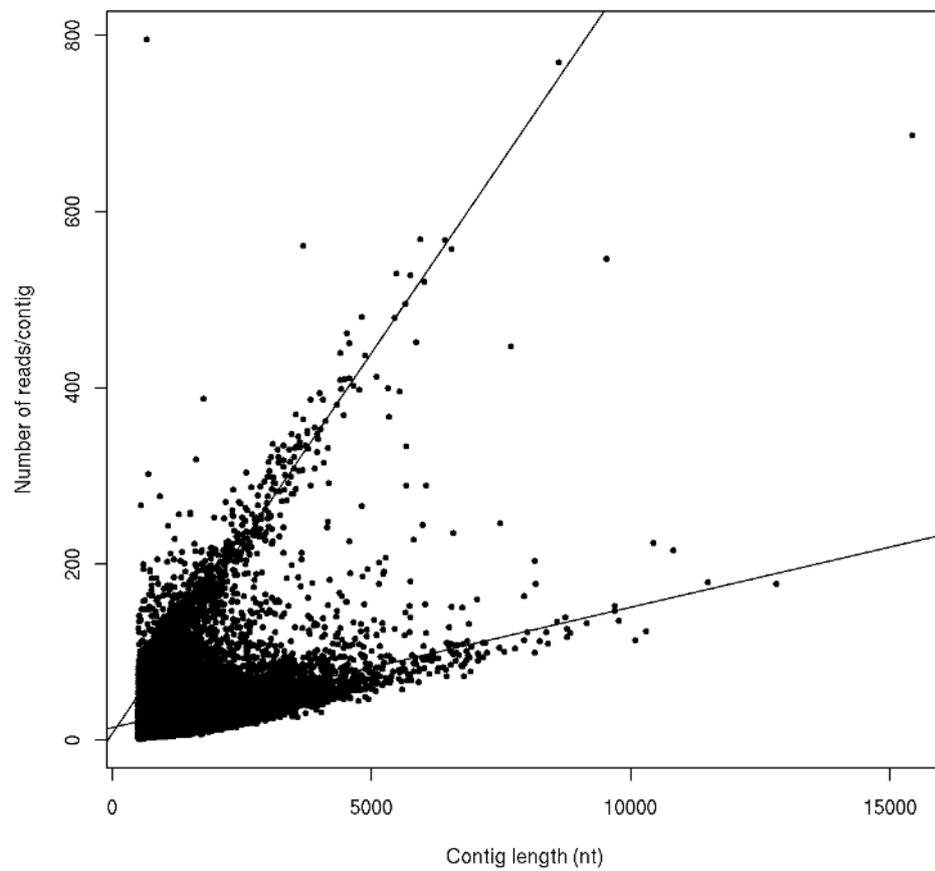
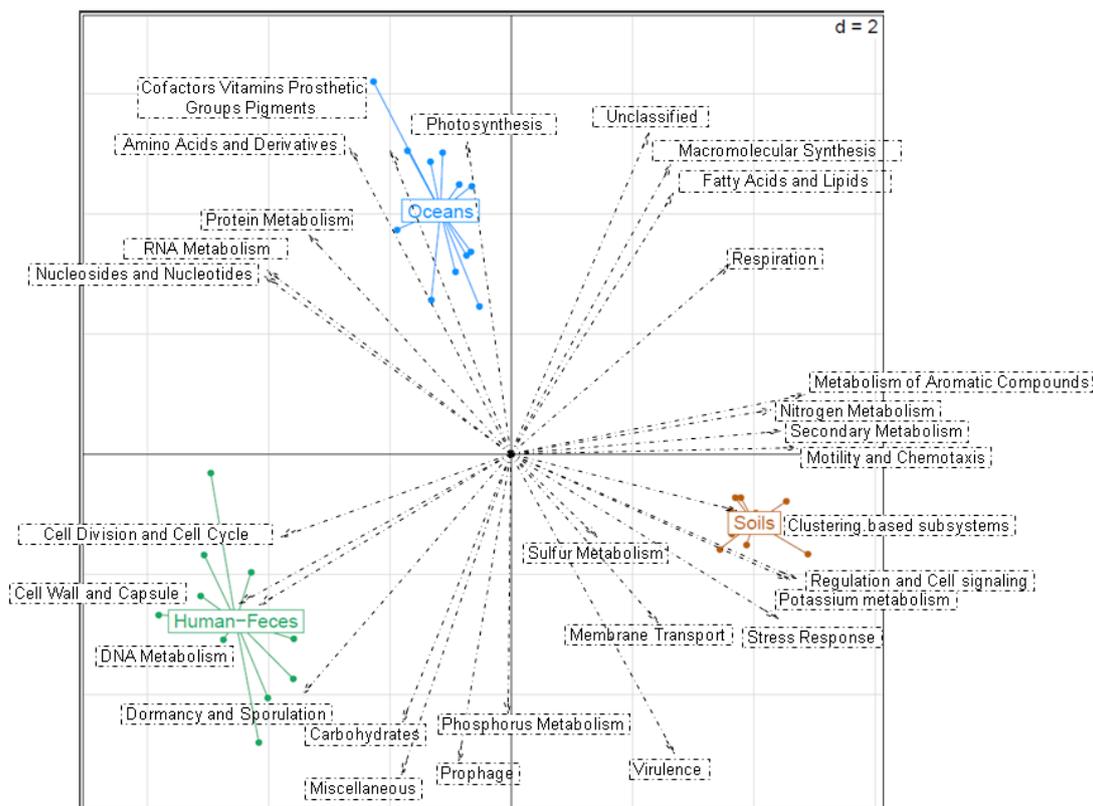


Figure 6



600

The submitted manuscript has been created in part by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

605