Improving the genetic representation of rare taxa within complex microbial communities using DNA normalization methods

DRAGANA GAGIC,* PAUL H. MACLEAN,† DONG LI,* GRAEME T. ATTWOOD* and CHRISTINA D. MOON*
*AgResearch Ltd, Grasslands Research Centre, Palmerston North, New Zealand, †AgResearch Ltd, Ruakura Research Centre, Hamilton, New Zealand

Abstract

Complex microbial communities typically contain a large number of low abundance species, which collectively, comprise a considerable proportion of the community. This ‘rare biosphere’ has been speculated to contain keystone species and act as a repository of genomic diversity to facilitate community adaptation. Many environmental microbes are currently resistant to cultivation, and can only be accessed via culture-independent approaches. To enhance our understanding of the role of the rare biosphere, we aimed to improve their metagenomic representation using DNA normalization methods, and assess normalization success via shotgun DNA sequencing. A synthetic metagenome was constructed from the genomic DNA of five bacterial species, pooled in a defined ratio spanning three orders of magnitude. The synthetic metagenome was fractionated and thermally renatured, allowing the most abundant sequences to hybridize. Double-stranded DNA was removed either by hydroxyapatite chromatography, or by a duplex-specific nuclease (DSN). The chromatographic method failed to enrich for the genomes present in low starting abundance, whereas the DSN method resulted in all genomes reaching near equimolar abundance. The representation of the rarest member was increased by approximately 450-fold. De novo assembly of the normalized metagenome enabled up to 18.0% of genes from the rarest organism to be assembled, in contrast to the un-normalized sample, where genes were not able to be assembled at the same sequencing depth. This study has demonstrated that the application of normalization methods to metagenomic samples is a powerful tool to enrich for sequences from rare taxa, which will shed further light on their ecological niches.

Keywords: DNA normalization, duplex-specific nuclease, high-throughput sequencing, hydroxyapatite chromatography, metagenomics, microbial community, rare biosphere

Received 4 March 2014; revision received 28 July 2014; accepted 13 August 2014

Introduction

Complex microbial communities are found in an extensive range of habitats on earth (Fierer & Lennon 2011; Lauber et al. 2013), and they typically contain a large number of genetically diverse, low abundance species. Collectively, these populations comprise a considerable proportion of the community, and they have been described as the ‘rare biosphere’ (Sogin et al. 2006). Although the size and importance of rare biospheres is still a topic of debate (Reeder & Knight 2009), the prevalence of rare biospheres is well documented (Pedros-Alio 2012), though their ecological and functional roles are not well understood. It has been hypothesized that the rare biosphere represents the products of historical adaptation, and is a valuable repository of genetic diversity (like a microbial seed bank) to facilitate community adaptation to environmental change (Lennon & Jones 2011; Caporaso et al. 2012). It has also been speculated that rare biosphere members may include keystone species – species that have a disproportionately large effect on their environment, relative to their abundance (Hajishengallis et al. 2012). The vast majority of environmental microbes are resistant to cultivation with current methods (Rappe & Giovannoni 2003), and this, in combination with their relatively low abundances, can pose significant challenges for understanding the ecological niches of rare biosphere members.

Currently, one powerful approach available to study the ‘rare biosphere’ is by metagenomic analysis, which is the study of genetic material recovered directly from
environmental samples (Handelsman et al. 1998). Metagenomics therefore allows the exploration of uncultivated taxa and uncovering of their metabolic potential. Theoretically, shotgun sequencing of the metagenome should reveal the roles and functions of rare community members. However, for highly complex communities, such as those that reside in gut, soil and aquatic environments, the depth of sequencing required to access rare members is significant, and major sequencing efforts to date have only enabled genome assembly from the most abundant community members (Hess et al. 2011; Iverson et al. 2012). Refinement of sequence data analyses using binning methods have enabled the genomes of rare taxa to be assembled, though, large volumes of sequence data are still required (Albertsen et al. 2013). Single cell genomics (Woyke et al. 2010) is a relatively recent strategy that will allow the rare biosphere to be accessed (Rinke et al. 2013); however, the technical demands of this approach are still beyond the reach of most standard laboratories.

An alternative approach to direct metagenome sequencing to improve the representation of rare biosphere is to apply DNA normalization methods to metagenomic DNA samples. DNA normalization methods are commonly applied to cDNA samples to facilitate the detection of rare transcripts, as the relative abundances of different transcript types can vary by orders of magnitude (Shagin et al. 2002). A schematic outlining the general normalization process is shown in Fig. 1. Traditionally, DNA normalization was achieved by hydroxyapatite (HAP) chromatography (Bernardi 1965; Sambrook & Russell 2001), though recently, the use of a thermostable duplex-specific nuclease (DSN) (Shagin et al. 2002) has become more common (Bogdanova et al. 2008). For both approaches, normalization is based on DNA renaturation kinetics, where the more abundant DNA sequences hybridize more rapidly than the rarer sequences. During HAP chromatography, the double-stranded and single-stranded DNA (dsDNA and ssDNA, respectively) molecules are physically separated by the differential charge interactions between the Ca$^{2+}$ ions on the surface of the HAP and the negatively charged phosphate backbone of the nucleic acids. For DSN-based DNA normalization, the dsDNA is enzymatically digested, leaving ssDNA from the lowly abundant DNA sequences intact.

HAP and DSN-based DNA normalization approaches have been used to normalize cDNA samples, and to remove repetitive sequences from genomic DNA to facilitate genome sequencing (Yuan et al. 2003; Bogdanova et al. 2008; Shagina et al. 2010). More recently, the advent of RNA-Seq to explore transcriptomes has driven the development of methods that enrich for mRNA by depleting the highly abundant rRNA from total RNA samples. Among rRNA depletion approaches, DSN and HAP chromatography have shown promising results in RNA samples from individual prokaryotes (Yi et al. 2011; Vandernoot et al. 2012), with a view to applying such methods to community total RNA samples (metatranscriptomics) (Carvalhais et al. 2012; Giannoukos et al. 2012).
Metatranscriptome normalization has potential to provide insight into rare biosphere member activities; however, the identification of transcripts from rare biosphere members initially requires reference metagenomic sequence to identify rare sequences.

Metagenome sequence reveals the functional potential of whole microbial communities, and provides reference genes and genomes to which metatranscriptome sequence data can be mapped. However, HAP and DSN-based normalization of metagenomic DNA samples has been very limited (Short & Mathur 1999), and critical assessment of the success of metagenome normalization has been lacking. As compared to traditional cDNA samples, where transcripts are of discrete lengths, from single organisms, and of relatively low complexity, the normalization of metagenomic DNA from complex microbial communities presents additional technical challenge. This is due to the greater sequence length heterogeneity and higher complexity of metagenomic DNA. Furthermore, microbial communities generally contain closely related organisms, and the hybridization of highly similar, but not identical, DNA sequences will result in the formation of heteroduplex molecules (that are not base paired fully along their lengths). How heteroduplexes behave during dsDNA removal by HAP chromatography and DSN treatments is not known, but if one sequence of the heteroduplex is rarer than the other, it may be eliminated in the dsDNA removal step. Metagenomes also contain many ‘repeat’ sequences (such as mobile genetic elements) flanked by unique sequences which can give rise to heteroduplexes, which will also influence the success of normalization.

In this study, we have applied previously established DSN and HAP chromatography normalization methods to a mock (synthetic) metagenome based on five gut-associated bacteria, and have critically assessed the efficiency of normalization by deep coverage shotgun sequencing of the resulting DNA samples, and determining changes in genome representation and genome coverage. De novo sequence assembly of the sequence data sets was also performed to determine the recovery of genetic information from rare community members. Such approaches will not only shed further light on the biology of the rare biosphere, but can significantly improve the efficiency of prospecting microbial communities for unique bioactivities that may have valuable industrial applications.

Materials and methods

Bacterial strains and growth conditions

The bacterial strains used in this study are listed in Table 1. *Escherichia coli* O127:H6 (strain E2348/69) and *E. coli* REL606 were grown in Luria–Bertani (LB) broth (Bertani 1952) at 37 °C with aeration. *Butyrivibrio proteoclasticus* B316T and *Prevotella ruminicola* 23 were cultivated anaerobically under a CO2 atmosphere in complete carbohydrate medium at 39 °C (Leedle & Hespell 1980). *Lactococcus lactis* IL1403 was grown in M17 broth (Merck,

| Strain                        | Genome size (Mbp) | %G+C content | Initial relative abundance* | Other features                                                                 | GenBank accession(s)       | Reference                  |
|-------------------------------|-------------------|--------------|------------------------------|-------------------------------------------------------------------------------|---------------------------|----------------------------|
| *Escherichia coli* O127:H6    | 5.07              | 50.7         | 100                          | Gram negative, one chromosome (4.97 Mb), two plasmids (98 kb and 6.1 kb), enteropathogenic, seven 16S rRNA genes | NC_011601–NC_011603       | Iguchi et al. (2009)       |
| Butyrivibrio proteoclasticus B316T | 4.40              | 39.3         | 100                          | Gram positive, two chromosomess (3.55 Mb and 302 kb), two megaplasmids (361 kb and 186 kb), rich glycobiome, six 16S rRNA genes | NC_014387–NC_014390       | Kelly et al. (2010)        |
| *Prevotella ruminicola* 23    | 3.62              | 47.7         | 10                           | Gram negative, rich glycobiome, one chromosome, four 16S rRNA genes            | NC_014033                 | Purushe et al. (2010)      |
| *Escherichia coli* B REL 606  | 4.63              | 50.8         | 10                           | Gram negative, one chromosome, seven 16S rRNA genes                           | NC_012967                 | Jeong et al. (2009)        |
| *Lactococcus lactis* subsp. lactis IL1403 | 2.37              | 35.3         | 1                            | Gram positive, one chromosome, lactic acid producer, six 16S rRNA genes        | NC_002662                 | Bolotin et al. (2001)      |

*Molar abundance in the synthetic metagenome sample.
Darmstadt, Germany) (Terzaghi & Sandine 1975) supplemented with 0.5% w/v glucose at 30 °C.

Genomic DNA extraction

High molecular weight genomic DNA was isolated from bacterial cultures using a modification of previously described methods (Stein et al. 1996; Liles et al. 2008), and is described in Appendix S1 (Supporting information).

Creation of synthetic metagenome and preparation for normalization

The synthetic metagenome DNA sample was created by mixing the genomic DNA from the five bacterial strains of interest in the molar ratio shown in Table 1. Approximately 35 µg of the synthetic metagenome DNA was digested to completion with PsiI (sample denoted as SM, see Fig. 2), and subsequently 20 µg of the digested DNA was ligated to a ‘lone linker’ in an excess of 100:1 (linker:DNA) molar ratio. The lone linker was generated by annealing LL-RIA (5’-pGAGATATTAGAATTCTACTC-3’) and LL-RIB (3’-TATAATCTTAAGATGAGp-3’) oligonucleotides (Ko et al. 1990) overnight at room temperature. Excess linker was removed by washing twice with 10 volumes of sterile water in a Vivaspin 100 micro-concentrator (Sartorius Stedim Biotech GmbH, Goettingen, Germany), and the DNA resuspended in 400 µL water. Using the LL-RIA oligonucleotide as a primer (0.4 µmol), the synthetic metagenome was PCR amplified by Platinum Taq DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA, USA), in the supplier’s buffer at 1× concentration, with 2.5 mM MgSO₄ and 0.8 µM dNTPs. Thermocycling, on a DNA Engine Peltier Thermal Cycler (Bio-Rad, Hercules, CA, USA), had an initial denaturation step for 2 min at 94 °C, followed by 25 cycles of denaturation (94 °C for 30 s), annealing (55 °C for 30 s) and extension (68 °C for 4 min). The extension step in the final cycle was extended to 7 min. The linker-amplified SM (SMLA) was purified

Fig. 2 Schematic overview of experimental and bioinformatic workflow.
using a QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).

**DSN-based DNA normalization**

A DSN-based DNA normalization was performed in triplicate on the SMLA sample as previously described (Shagina et al. 2010). A hybridization mixture contained approximately 2.75 μg amplified synthetic metagenome mixed with 11 μL of 4X hybridization buffer [200 mM HEPES (pH 7.5), 2 mM NaCl, 0.8 mM EDTA] (Shagina et al. 2010) in nuclease free water up to a total reaction volume of 44 μL. The reaction mixture was aliquoted into 11 × 0.2 mL PCR tubes, and each aliquot was overlaid with mineral oil. Denaturation of DNA was performed at 98 °C for 3 min, followed by renaturation at 68 °C for 5 h, in a DNA Engine Peltier Thermal Cycler (Bio-Rad). Reactions were incubated at 65 °C prewarmed 10× DSN Master buffer (Evrogen, Moscow, Russia) was added to each reaction mixture, taking care to maintain the 68 °C hybridization temperature. To ten of the 11 tubes, 1/8 U of DSN (Evrogen) was added. Reactions were incubated at 65 °C for 20 min, and then, DSN was inactivated by addition of EDTA to a final concentration of 2 mM. The remaining DNA was amplified using primer LL-RIA and Platinum PCR SuperMix High Fidelity (Invitrogen). PCR products were purified by phenol–chloroform–isoamyl alcohol (25:24:1), precipitated with ethanol and diluted in nuclease free water to a final DNA concentration of 100 ng/μL. Five rounds of DSN-based DNA normalization were performed and the PCR fragment size distribution after each round was assessed by agarose gel electrophoresis.

**HAP-based DNA normalization**

HAP chromatography-based normalization was also performed in triplicate on the SMLA sample. SMLA DNA was denatured and hybridized under the conditions described for DSN-based normalization except that approximately 10 μg DNA was used per reaction and the reaction mixture up-scaled accordingly. After renaturation, the DNA mixture was added to 60 °C prewarmed 0.12 mM phosphate buffer (pH 6.8) to a total volume of 0.5 mL and the mixture was incubated for 15 min. HAP chromatography to separate the dsDNA from the ssDNA was performed based on the protocol of (Andrews-Pfannkoch et al. 2010). Hydration of DNA grade Bio-Gel HTP hydroxyapatite (Bio-Rad) and preparation of the jacketed 0.7 × 15 cm Econo-Column (Bio-Rad) were performed according to the manufacturer’s recommendations. The inside of the column was coated with 1 mL of Sigmacote (Sigma-Aldrich, St. Louis, MO, USA) and allowed to air dry to prevent nonspecific nucleic acid binding. Throughout the fractionation, all buffers and hydrated HAP were maintained at 60 °C. The temperature of the column was maintained at 60 °C by externally insulating the column with cotton wool, and circulating 60 °C water from a water bath through the jacket. The column was rinsed twice with 7 mL of 0.12 mM phosphate buffer, then 2 mL of preprepared, resuspended HAP was gradually added to the bottom of the column and allowed to settle for 30 min. The warm (65 °C) hybridization mixture was quickly added to the HAP to avoid cooling, and DNA was allowed to bind for 30 min at 60 °C. The initial flow-through was discarded, and the column was sequentially loaded and eluted with 0.18, 0.2, 0.4 and 1 mM prewarmed phosphate buffers, with eluates collected separately after each buffer addition. The 0.18 and 0.2 mM eluates (containing ssDNA), and the 0.4 and 1 mM eluates (containing dsDNA) were each pooled together and desalted using Amicon Ultra-4 centrifugal filter devices with 30 kDa molecular weight cutoff (Millipore). The resulting ssDNA was ethanol precipitated and resuspended in 50 μL of nuclease free water. The ssDNA and dsDNA fractions (as controls) were amplified using the LL-RIA primer as previously described, and purified products were used for the next round of DNA normalization. Five rounds of HAP chromatographic DNA normalization were performed, and the size distribution of the PCR products from each round were visualized by agarose gel electrophoresis.

**PCR denaturing gradient gel electrophoresis**

To assess shifts in the abundance of the 16S rRNA genes within the synthetic metagenome samples, the V6 regions of the 16S rRNA genes were amplified with primers EUB 933f-GC-clamp (5’-GCACAAGCCTGTGAGCAGTGTGG-3’) (Iwamoto et al. 2000) and EUB1053r (5’-AGCTGACGACAGCCATGCA-3’) and PCR denaturing gradient gel electrophoresis (DGGE) was performed as described in Appendix S1 (Supporting information).

**Synthetic metagenome sequencing**

The original PsiI-digested synthetic metagenomic DNA (SM sample), the lone linker-amplified PsiI-digested synthetic metagenomic DNA (SMLA sample), and the PCR amplicons generated over the five rounds of normalization using the DSN (DSN_R1–R5) and HAP-based (HAP_R1–R5) normalization methods (see Fig. 2), were sequenced at MACROGEN (Seoul, Korea) using the Illumina HiSeq2000 platform. Barcoded sequencing samples were prepared using a Nextera XT DNA sample preparation kit (Illumina Inc., San Diego, CA, USA). In total, 36 samples were run on four lanes of a HiSeq2000 flowcell, and the three replicates of each sample type were distributed
across the four lanes to minimise potential lane to lane variation effects. Paired end sequencing was performed with 100 bp read lengths.

**Sequence analysis**

Overall genome similarities were determined using DNA–DNA hybridization (DDH) values in the Genometo-Genome Distance Calculator version 2 (Auch et al. 2010) using BLAST+ as the alignment method.

The sequence reads were checked for quality and nucleotide composition biases using FASTQC (Fulweiler et al. 2012) and the FASTX toolkit [fastx_quality_stats function; (Bartual et al. 2010)]. CUTADAPT (Martin 2011) was used to remove the sequencing adaptors, lone linker sequences (using an overlap of six base pairs), and the regions of low quality (<Q30) sequence.

The sequence reads of the SM, SMLA and all normalized samples (Fig. 2) were mapped against in silico PsiI-digested genomes of the five bacterial strains comprising the synthetic metagenome using Burrows–Wheeler Aligner (BWA) v0.6.2-r126 (Li & Durbin 2009) with default parameters. Read pairs for which at least one of the reads uniquely mapped to a PsiI fragment, or both paired reads mapped to the same PsiI fragment, were counted per read pair using a custom perl script (Appendix S2, Supporting information) and enumerated for each PsiI fragment. The counts for these fragments were collated using MICROSOFT EXCEL’s vlookup function. The counts for the fragments were summed for each organism (fragments from both the main chromosome and any additional replicons). The average read coverage per PsiI fragment was visualized using CGView (Stothard & Wis- hart 2005), where the log of the fold coverage of the PsiI fragment was displayed as a proportion of that of the fragment with maximum average fold coverage for all PsiI fragments in the genome of the same strain. Differences between the accumulated counts for each organism and fragment sizes were tested using the two-sided Student’s t-test with unequal variance implemented in MICROSOFT EXCEL. Differences between observed count data and theoretical distributions were tested using a one sample Student’s t-test in R (R Core Team 2013) with the theoretical values as the null hypothesis.

Sequence reads for each sample were assembled into contigs using the VELVET de novo assembler version 1.2.09 (Zerbino & Birney 2008), and optimized using the VELVET-optimiser perl script (Gladman & Seemann 2012) using kmer sizes between 55 and 69, optimising for N50 length. Both the open reading frames (ORFs) and the PsiI fragments of the reference metagenomic sequence were searched against the assembled contigs using the BLASTN program of the BLAST+ suite (Dethlefsen et al. 2008). The top high-scoring segment pair (HSP) for each coding sequence ORF and PsiI fragment was parsed from the BLASTN results using an in-house BLAST parser (Appendix S3, Supporting information). Assemblies were also performed on sequence reads from each sample that were normalized in silico using Diginorm (Brown et al. 2012), with a single pass approach, kmer length of 21, and coverage cut-off of 10. The resulting normalized read pairs were assembled into contigs using the VELVET de novo assembler, and analysed as described above.

**Results and discussion**

**Experimental normalization of the synthetic metagenome**

Within natural microbial communities, significant differences in the abundance of different taxa, and genetic similarity between taxa, are common features. We therefore tested two experimental normalization methods, employing HAP chromatography and DSN, on a synthetic metagenomic DNA sample comprised of genomic DNAs pooled from the five bacterial strains listed in Table 1. These strains, which are all associated with gut environments or food production, were specifically chosen for their diverse genomic characteristics, such as genome size, G+C% content, presence/absence of multi-copy replicons and the availability of complete genome sequences. The synthetic metagenome was constructed by pooling high molecular weight DNA from the five strains in a defined molar ratio of 1000:100:10:10:1 for Escherichia coli E2348/69; Butyribri vibrio proteoclasticus B315; Prevotella ruminicola 23; E. coli REL606; Lactococcus lactis IL1403. Such pooling was performed to allow the investigation of (a) the effective range of normalization by these methods (testing over three orders of magnitude in abundance) and (b) whether an abundant genome can be normalized relative to a less abundant but very closely related genome for example E. coli strains E2348/69 and REL606. To determine the degree of genetic difference between the strains, pairwise genome-to-genome distance calculations were undertaken (Table S1, Supporting information). The two most closely related strains, E2348/69 and REL606, had a DDH value of 74.9%, which is in contrast to all other pairwise comparisons where values ranged from 12.6% to 22.5%.

In contrast to cDNA normalization, where cDNA molecules are generally of lengths that are able to be PCR amplified using standard conditions, it was necessary to fragment the synthetic metagenomic DNA to obtain such a PCR-amplifiable sample (SM sample). Here, restriction digestion was preferred over shearing methods to minimise the occurrence of heteroduplex molecules forming during hybridization due to the mis-alignment of fragment ends. Restriction endonuclease
PsiI was chosen as it generated blunt-ended fragments to facilitate downstream linker attachment, and for the genomes selected, it yielded a relatively high proportion of fragments up to 5 kb in size, suitable for PCR amplification. Lone linkers (Ko et al. 1990) were ligated to the fragment ends, and the synthetic metagenome was amplified (SMLA sample) to generate a pool of material to test using both normalization procedures. Notably, the amplification of the synthetic metagenome resulted in the enrichment of smaller PsiI fragments (the majority were less than 4 kb), whereas the vast bulk of fragments in the PsiI-digested sample were greater than 3 kb. It is recognized that metagenomic DNA digestion by a single enzyme will affect genome representation (as the large restriction fragments will be under-represented after amplification) and that numerous genes will be cleaved (which will affect downstream functional screening of normalized samples). Furthermore, for any metagenome sequencing, assembly will be limited as the restricted fragments do not overlap. To minimise these effects, in practice, the normalization process should be repeated using different restriction enzymes in parallel, and pooling the resulting collection of normalized fragments to create a set of overlapping fragments. However, due to resource constraints and as our primary focus was to identify efficient methods for metagenome normalization, multiple enzymes were not applied in this study.

The SMLA sample was thermally denatured, and stringently hybridized under controlled conditions, then subjected to dsDNA removal either by HAP chromatography or by DSN digestion. Prior to normalising the SMLA sample, conditions for HAP chromatography and DSN digestion were tested using control mixtures of circular ssDNA, linear ssDNA and dsDNA samples, which confirmed that they could adequately separate or remove the dsDNA from the ssDNA (data not shown). The normalization methods were applied to the SMLA sample and the resulting ssDNA enriched samples were amplified via the lone linker sequence. This process was repeated for five cycles, removing a sample after each round to assess the extent of normalization throughout the procedure.

**Initial assessment of experimental normalization**

Assessment of normalization of the synthetic metagenome was performed in two steps – first by determining the relative abundance of a single locus (16S rRNA gene) using PCR-DGGE and then by assessing the coverage across each of the genomes of the syn-

---

**Fig. 3** PCR-DGGE of the 16S rRNA gene V6 region of (a) hydroxyapatite (HAP) chromatography, and (b) DSN normalized samples. Lanes R1–R5 refer to the normalization cycle number, and banding patterns for triplicate samples are shown; lanes P show pooled amplicons from individual genomic DNA samples in equimolar abundance; lanes S show amplicons from the SM DNA, and lanes SA show amplicons from the lone linker-amplified SMLA DNA. Normalization rounds where amplicons for each strain are at their greatest relative abundance are shown in ovals.
thetic metagenome via shotgun sequencing with deep coverage (200×). PCR-DGGE based on the phylogenetically informative V6 region of the bacterial 16S rRNA gene (Yu & Morrison 2004) was performed on all samples (Fig. 3). This provided an initial indication of the relative successes of the normalization methods. Normalization by both methods resulted in substantial enrichment of products from the rarer strains (IL1403, 23 and REL606; Fig. 3), though the number of rounds of normalization that maximum enrichment was observed varied by both strain and normalization method. The relative depletion of products from the most abundant strain, E2348/69, was most apparent by the DSN method in the fourth and fifth rounds while no apparent depletion was observed by the HAP method. Enrichment of products from the rarest strain, IL1403, was prominent in the second and third rounds of DSN normalization, and in the fifth round of HAP normalization, suggesting that enzymatic normalization was effective more quickly than HAP normalization. For the second most abundant strain, B316, no depletion of products was observed by either method. However, for strains 23 and REL606, which were present at the same starting abundance (and the second rarest strains), 23 showed an increase in abundance throughout the experiment for both methods, whereas REL606 bands were at their greatest intensity at rounds two and three, but depleted in subsequent rounds of normalization. Overall, shifts in the abundances of 16S rRNA gene amplicons were observed throughout both normalization processes, but as 16S rRNA genes are only a single locus, and typically present in multiple copies in bacterial genomes (Table 1), metagenome-wide assessment was required to fully assess the extent of normalization.

**Mapped sequence analysis of normalized metagenome samples**

Assessment of normalization across the entire synthetic metagenome, and over the five rounds (R1–R5) of normalization, was performed by shotgun sequencing the normalized samples (HAP_R1–R5 and DSN_R1–R5) and the non-normalized samples (SM and SMLA). Each sample was sequenced to a depth of approximately 200× based on the nonredundant (nr) synthetic metagenome (estimated to be approximately 20.1 Mb), using 100 bp paired end sequencing. An average of 23.05 and 24.49 M

![Fig. 4 Proportion of each genome in the synthetic metagenome and hydroxyapatite (HAP) and DSN chromatography normalized samples, represented by mapped shotgun sequence read pairs. Samples are SM, synthetic metagenome; SMLA, synthetic metagenome linker-amplified; HAP_R1-R5, HAP chromatography normalized samples, cycles 1–5; DSN_R1-R5, DSN normalized samples, cycles 1–5; Theoretical, the expected representation of each genome if present in equimolar ratio. Standard errors of the mean for each proportion are depicted by error bars. After one round of normalization, differences in the relative genome representation of each of the HAP_R1 and DSN_R1 samples, compared to the un-normalized SMLA sample (comparisons denoted by the connecting lines), were determined using the two-sided Student’s t-test with unequal variance. Significant differences in mapped read abundance per genome (P ≤ 0.01) are denoted by asterisks. For HAP_R1, the proportion means for IL1403 and 23 (for which, bars are not visible) were significantly different to the un-normalized sample.](image-url)
reads were obtained for the normalized HAP and DSN samples, respectively, and 22.75 M for the SM, and 22.21 M for the SMLA samples (Table S2, Supporting information). Sequence reads were mapped to the reference genome sequences to a) indicate the level of normalization at the genome level and b) highlight any areas of under- or overrepresentation and reveal any genomic features which may be difficult to normalize. It was anticipated that unique sequences adjacent to repeat sequences may fall in this category.

The mapped sequence read data are presented in Table S2 (Supporting information). For the SM sample, sequence reads were mapped to the reference genomes that had been digested by PsiI in silico, was the highest at 86.7%. However, after lone linker amplification (sample SMLA), the mapping rate dropped slightly, but significantly (P-value = 0.02 via Student’s t-test), to 85.0%, probably due to the introduction of sequence errors during PCR. During normalization, the normalized samples exhibited decreases in mapping rates from 81.0% to 70.7% for HAP chromatography normalized, and 86.4% to 69.3% for DSN normalized, over the five rounds of normalization. This is probably to reflect the accumulation of PCR errors over rounds of normalization. The extent of the reduction in mapping rate between subsequent rounds of DSN normalization approximately doubled each round (Table S2, Supporting information), whereas this trend was not apparent for the HAP samples. DSN_R1–R3 had a considerably greater proportion of reads mapped (86.4%, 85.2% and 82.9%, respectively), as compared to their corresponding HAP normalized samples (81.0%, 77.4% and 73.2% for R1–R3-3). This indicates that the DSN approach maintained reasonably high sequence integrity within the earlier rounds of normalization.

During mapping, occasionally reads from a single read pair mapped to two different PsiI fragments. These may represent a small degree of partial PsiI digestion of the synthetic metagenome. Furthermore, read pairs that did not map, or could not be mapped uniquely, to a PsiI fragment, were not considered further.

*Genome representation throughout normalization processes*

The representation of each genome throughout the normalization experiment, was determined *via* the proportion of mapped read pairs to each genome, and compared to the non-normalized samples (Fig. 4 and Table S2, Supporting information). Sequence coverage for the SM sample was consistent with the theoretical prediction based on the starting ratios of the pooled DNA, apart from *E. coli* REL606, where the expected proportion of reads was 0.84%; however, 3.91% of reads were observed in the SM sample (Table S3, Supporting information). This discrepancy may have been due to the accuracy of the REL606 DNA quantification, affecting the actual amount of REL606 DNA present within the SM sample.

For the HAP method, the proportion of reads mapped to each genome did not differ significantly compared to the SMLA sample, nor among all five cycles of normalization (Fig. 4). This was unexpected given the clear shifts in 16S rRNA gene abundance observed by DGGE analysis (Fig. 3), and the reason for this is not known. However, for the DSN normalized samples, there was a marked shift in the representation of the genomes after only a single round of normalization (see DSN_R1 compared to SMLA in Fig. 4). For example, the abundance of reads for E2348/69 was significantly decreased (from 86.2% to 10.2%; P-value = 0.0013), while for the rarest member, IL1403, representation increased from 0.07% to 1.48% (P-value = 0.0019), which is a greater than 20-fold increase after the first cycle of normalization. Subsequent rounds of normalization resulted in further shifts in genome representation (Fig. 4 and Table S2, Supporting information). After five rounds of DSN normalization, each genome represented approximately 6–42% of the total synthetic metagenome (Table S3, Supporting information).

Taking into account the differences in genome sizes, the mapped sequence proportions were compared to the theoretical distributions expected for the fully normalized metagenome (with each genome present in equimolar abundance; Table S3, Supporting information). For the least abundant strain, IL1403, the initial abundance of 0.04% of the mapped sequence reads increased to 18.32% after five rounds of DSN normalization, which represents over a 450-fold increase to constitute nearly one-fifth of the metagenome. Similarly, strain 23 displayed a substantial increase in representation (from 0.55% to 11.87%). In contrast, the representation of the most abundant strain, E2348/69, had decreased from 90.05% to 21.76%. Unexpectedly, B316<sup>T</sup> representation was substantially enriched from 5.45% to 42.25%, whereas an equimolar proportion should have been 21.93%. The reason for this degree of enrichment is unclear. REL606 sequences (3.91%) did not enrich to the extent predicted (23.1%) if the metagenome had been fully normalized (Fig. 4), reaching only 5.81% representation. The probably factor that contributed to this is REL606’s high sequence similarity with the closely related E2348/69. The highly similar regions within these genomes would probably promote the formation of REL606 – E2348/69 heteroduplexes during the hybridization step, and removal of such heteroduplexes by DSN digestion would further eliminate the rarer REL606 fragments. Despite not reaching the theoretical proportions for fully normal-
ized samples, the REL606 sequences increased in abundance by greater than threefold, while elimination of the E2348/69 sequences was greater than fourfold during the normalization procedure. These results indicate that DSN normalization was far more effective than HAP chromatography in normalising the synthetic metagenome sample; however, the degree of normalization of each genome was not uniform. This result is not unexpected given the complexity of the synthetic metagenome sample, which is comprised of organisms with a wide range of genetic features (Table 1).

**Genome coverage after normalization**

To assess the relative evenness of coverage for each genome during the course of the normalization processes, the average fold coverage of each PsiI fragment was determined for each genome, and the data for IL1403 is represented in Fig. 5. In the un-amplified SM sample, the theoretical depth of sequence coverage of IL1403 was only approximately 2×; thus, coverage of the IL1403 genome was fairly uneven (Fig. 5, SM ring). Furthermore, PCR amplification using the lone linker primer introduces considerable fragment length bias, and the evenness of genome representation becomes more sparse (Fig. 5, SMLA ring).

HAP chromatography-based normalization resulted in very poor evenness of genome coverage, even after five cycles of normalization (Fig. 5a, HAP_R1 to R5 rings). In contrast, by DSN normalization (Fig. 5b, DSN_R1 to R3 rings), the fold coverage (R1 = 28×, R3 = 138×, R5 = 379×) and general evenness of genome coverage over rounds of normalization increased. Genome coverage analysis for the other four genomes showed similar trends as for IL1403, where HAP chromatography-based normalization resulted in inconsistent genome coverage, whereas the DSN method achieved even genome coverage (see Fig. S1, Supporting information).

An unexpected result was that very large PsiI fragments (up to 63.5 kb, for E2348/69 in Fig. S1, Supporting information) were well represented after DSN normalization.

**Table 2** Number of ORFs detected in the *de novo* assembled metagenome samples before and after DSN normalization

| Strain                          | Initial relative abundance | Total ORFs | Assembled ORFs (%) |
|---------------------------------|---------------------------|------------|-------------------|
|                                 |                           |            | SM.1 DSN_R5.1     | DSN_R5.1 (Diginorm) |
| *Escherichia coli* E2348/69    | 1000                      | 4642       | 85.5              | 25.6                 |
| *Butyrivibrio proteoclasticus* B316 | 100                       | 3811       | 72.1              | 19.4                 |
| *E. coli* REL606                | 10                        | 4290       | 13.0              | 8.3                  |
| *Prevotella ruminicola* 23      | 10                        | 2763       | 0.0               | 7.6                  |
| *Lactococcus lactis* IL1403    | 1                         | 2266       | 0.0               | 13.5                 |

1 Molar abundance in the synthetic metagenome sample.
2 Proportion of ORFs that exhibited ≥99% sequence identity over ≥95% of the ORF length.
normalization. Such fragments are probably to be amplified with extremely low efficiency under the PCR conditions used, and therefore, are expected to be relatively rare in the initial SMLA sample. Sequence read mapping to the fragment at R5 revealed some unevenness of coverage (as compared to coverage in the SM sample), suggesting that localized amplifications of the fragment had occurred, potentially due to mis-priming of the lone linker primer during amplification.

To determine the effectiveness of DSN normalization in reducing the representation of repeated sequences, such as rRNA operons and insertion sequence (IS) elements, we examined the relative abundances of 16S rRNA gene sequence reads within each genome before and after normalization. The 16S rRNA gene frequency (Table 1) decreased to levels that approached one copy per genome (see Fig. S2, Supporting information), as expected for complete normalization.

De novo assembly of normalized metagene data sets

The assembly of genomes from members of the rare biosphere would shed considerable light on their functions, thus, to assess the degree of genome reassembly from the normalized and non-normalized synthetic metagenomes, de novo assemblies were performed. As the synthetic metagenome was initially digested with PsiI, the proportion of PsiI fragments reassembled was first examined (Table S4, Supporting information). Given the average depth of coverage (200×) of the nr SM and the large discrepancies in relative abundances of the genomes in the SM sample, only 34% of the PsiI fragments could be assembled ≥99% sequence identity for ≥95% of the length of the fragment (Table S4, Supporting information). Upon lone linker amplification, which would generally bias against long PsiI fragments, the PsiI fragment assembly rate dropped slightly to 32% for the SMLA sample. After five rounds of DSN normalization, the relative abundance of each genome was more even, and 25% of the PsiI fragments could be reconstructed. The inevitable introduction of PCR errors throughout the normalization, sequencing errors and potential mis-priming of the sequence with the lone linker primer, probably contributed to the lower assembly of contiguous sequences. However, for application to a real metagenome, one should perform normalization in parallel with multiple restriction enzymes to generate sets of overlapping sequences that would aid assembly.

For biodiscovery applications, we wanted to see how many genes could be reassembled within the normalized data sets, particularly from the rare organisms (Table 2). In the SM sample, the majority of genes reassembled were from the most abundant organisms (85.5% for E2348/69, and 72.1% for B316). Representation of the rare genomes (23 and IL1403) was so low that genes did not reassemble (Table 2). After five rounds of DSN normalization, genes from the initially rare members were recovered (7.6% for 23, and 13.5% for IL1403). Correspondingly, the frequencies of genes from the most abundant organisms were substantially decreased (25.6% for E2348/69, and 19.4% for B316). For REL606, the second rarest organism, the number of ORFs assembled decreased (13.0–8.3%), which is probably related to its normalization dynamics in the presence of the closely related E2348/69.

The VELVET assembler is primarily designed for single genome assembly; thus, assemblies were also generated by first normalising the sequence reads in silico using Diginorm (Brown et al. 2012) to generate more uniform coverage of the genomes prior to assembly with VELVET. Diginorm resulted in a substantial reduction of read pairs, with approximately 3–12% of read pairs retained. Assembly of the digitally normalized data, as compared to the direct VELVET assemblies, is shown in Table S4 (Supporting information). Curiously, in the original metagenome data set (SM.1), the total assembly size increased (from 9.3 to 12.9 Mb); however, the percentage of PsiI fragments that were reconstructed had decreased considerably (from 34% to 25%). For the DSN normalized data set (DSN_R5.1), the total assembly size also increased (from 11.9 to 14.2 Mb), yet the percentage of PsiI fragments reconstructed increased from 25% to 34% (Table S4, Supporting information). In the DSN normalized data set, the percentage of genes assembled had increased for all strains (Table 2), and most strikingly, for IL1403, the increase was from 13.5% to 18.0%.

Complex microbial community samples

We have here demonstrated the effectiveness of DSN normalization on a synthetic metagenomic DNA sample. How this will translate to an actual microbial community sample, which is comprised of thousands to millions of strains of varying genetic relatedness, is of significant interest. After genomes are fragmented, normalization will act at a sequence, rather than genome, level. Thus, it is predicted that abundant sequences, such as those from highly abundant organisms, conserved repetitive sequence elements and highly conserved homologous genes (for example, the core genome sequences of pan-genomes from species-complexes (Bentley 2009) will be depleted first). Enrichment will probably occur for rare sequences, such as those from rare members, or strain-specific accessory genes (Bentley 2009), that are genetically disparate to the majority sequences within the metagenome. We intend to follow up this study by applying DSN-based normalization to metagenomic
DNA from complex microbial communities, and evaluates its effectiveness in enriching for their rare biospheres and rare genes.

Summary

The DSBN-based normalization of a synthetic metagenome derived from five bacterial strains resulted in significantly increased representation of the sequences from the rarest members of the community. Furthermore, normalization resulted in the ability to assemble genes from the rare members that could not be detected prior to normalization using the same depth of sequencing. This study has demonstrated that the application of normalization methods to metagenomic samples is a useful tool to substantially reduce the number of abundant sequences from common members of microbial communities, and enrich for sequences from the rare taxa, which will improve our understanding of their functions, offer new insights into their ecological roles, as well as aid in the discovery of new bioactivities with industrial applications.

Acknowledgements

We wish to thank Richard Lenski, Adrian Cookson and Bill Kelly for the provision of bacterial strains, Vanessa Cave for statistical guidance, and Eric Altermann, Janine Kamke and two anonymous reviewers for critically reviewing the manuscript. We gratefully acknowledge funding from the AgResearch Capability Emerging Leader Fund, contract A15117.

References

Albertsen M, Hugenholtz P, Skarszewski A et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology, 31, 533–538.

Andrews-Pfannkoch C, Fadrosh DW, Thorpe J, Williamson SJ (2010) Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. Applied and Environment Microbiology, 76, 5039–5045.

Auch AF, von Jan M, Klenk HP, Goker M (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Standards in Genomic Sciences, 2, 117–134.

Bartual SC, Otero JM, Garcia-Doval C et al. (2010) Structure of the bacteriophage T4 long tail fiber receptor-binding tip. Proceedings of the National Academy of Sciences of the USA, 107, 20287–20292.

Bentley S (2009) Sequencing the species pan-genome. Applied and Environment Microbiology, 75, 258–259.

Bernardi G (1965) Chromatography of nucleic acids on hydroxyapatite. Nature, 206, 779–783.

Berti G (1952) Studies on Lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli. Journal of Bacteriology, 62, 293–300.

Bogdanova EA, Shagin DA, Lukyanov SA (2008) Normalization of full-length enriched cDNA. Molecular BioSystems, 4, 205–212.

Bolotin A, Wincker P, Mauger S et al. (2001) The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1400. Genome Research, 11, 731–753.

Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH (2012) A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv:1203.4802v2.

Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA (2012) The Western English Channel contains a persistent microbial seed bank. ISME Journal, 6, 1089–1093.

Carvalhais LC, Dennis PG, Tyson GW, Schenk PM (2012) Application of metatranscriptomics to soil environments. Journal of Microbial Methods, 91, 246–251.

Defelice L, Huse S, Sorgen ML, Reisman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biology, 6, e280.

Fierer N, Lennon JT (2011) The generation and maintenance of diversity in microbial communities. American Journal of Botany, 98, 439–448.

Fulweiler RW, Rabalais NN, Heiskenan AS (2012) The eutrophication commandments. Marine Pollution Bulletin, 64, 1997–1999.

Giannoukos G, Ciulla DM, Huang K et al. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biology, 13, R23.

Gladman S, Seemann T (2012) VelvetOptimiser. Available from http://www.vicbioinformatics.com/software/velvetoptimiser.shtml.

Hajishengallis G, Darveau RP, Curtis MA (2012) The keystone-pathogen hypothesis. Nature Reviews Microbiology, 10, 717–725.

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. Chemistry & Biology, 5, R245–R249.

Hess M, Sczyzrba A, Egam R et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science, 331, 463–467.

Iguchi A, Thomson NR, Ogura Y et al. (2009) Complete genome sequence and comparative genome analysis of enteropathogenic Escherichia coli O127:H6 strain E2348/69. Journal of Bacteriology, 191, 347–354.

Iversen V, Morris RM, Frazar CD et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science, 335, 587–590.

Iwamoto T, Tani K, Nakamura K et al. (2000) Monitoring impact of in situ biostimulation treatment on groundwater bacterial community by DGGE. FEMS Microbiology Ecology, 32, 129–141.

Jong H, Barbe V, Lee CH et al. (2009) Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). Journal of Molecular Biology, 394, 644–652.

Kelly WJ, Leahy SC, Altermann E et al. (2010) The glyco biome of the ruminant bacterium Butyrivibrio proteoclasticus B316(T) highlights adaptation to a polysaccharide-rich environment. PLoS ONE, 5, e11942.

Ko MS, Ko SB, Takahashi N, Nishiguchi K, Abe K (1990) Unbiased amplification of a highly complex mixture of DNA fragments by ‘lone linker’-tagged PCR. Nucleic Acids Research, 18, 4293–4294.

Lauber CL, Ramirez KS, Aanderud Z, Lennon J, Fierer N (2013) Temporal variability in soil microbial communities across land-use types. ISME Journal, 7, 1641–1650.

Leedle JA, Hessell RB (1980) Differential carbohydrate media and anaerobic replica plating techniques in delineating carbohydrate-utilizing subgroups in rumin bacterial populations. Applied and Environment Microbiology, 39, 709–719.

Lennon JT, Jones SE (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. Nature Reviews Microbiology, 9, 119–130.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25, 1754–1760.

Liles MR, Williamson LL, Rodbrunner J et al. (2008) Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. Applied and Environment Microbiology, 74, 3302–3305.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal, 17, 10–12.
Pedros-Alio C (2012) The rare bacterial biosphere. Annual Review of Marine Science, 4, 449–466.

Purusha J, Fouts DE, Morrison M et al. (2010) Comparative genome analysis of Prevotella ruminicola and Prevotella bryantii: insights into their environmental niche. Microbial Ecology, 60, 721–729.

R Core Team (2013) R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annual Review of Microbiology, 57, 369–394.

Reeder J, Knight R (2009) The ‘rare biosphere’: a reality check. Nature Methods, 6, 636–637.

Rinke C, Schwientek P, Sczyrba A et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature, 499, 431–437.

Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor NY.

Shagina I, Bogdanova E, Mamedov IZ et al. (2010) Normalization of genomic DNA using duplex-specific nuclease. BioTechniques, 48, 455–459.

Short JM, Mathur EJ (1999) Constructing a genomic deoxyribonucleic acid library from an exogenous sample: obtaining exogenous genomic sample, amplifying genomic sample, calibrating genomic library; for screening libraries of mixed microorganism samples. US Patent 6001574 A.

Sogin ML, Morrison HG, Huber JA et al. (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proceedings of the National Academy of Sciences of the USA, 103, 12115–12120.

Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of Bacteriology, 178, 591–599.

Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. Bioinformatics, 21, 537–539.

Terzaghi BE, Sandine WE (1975) Improved medium for lactic streptococci and their bacteriophages. Applied Microbiology, 29, 807–813.

Vandernoot VA, Langevin SA, Solberg OD et al. (2012) cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. BioTechniques, 53, 373–380.

Woyke T, Tighe D, MAVROMATIS K et al. (2010) One bacterial cell, one complete genome. PLoS ONE, 5, e10314.

Yu Z, Morrison M (2004) Comparisons of different hypervariable regions of 16S rDNA genes for use in fingerprinting of microbial communities by PCR-denaturing gradient gel electrophoresis. Applied and Environment Microbiology, 70, 4800–4806.

Yuan Y, SanMiguel PJ, Bennetzen JL (2003) High-Cot sequence analysis of the maize genome. The Plant Journal, 34, 249–255.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18, 821–829.

C.M. and D.G. designed the research, C.M. and G.A. obtained funding for the study, D.G. and D.L. conducted the experimental work, P.M. and C.M. analysed data, D.G., P.M. and C.M. wrote the manuscript.

Data accessibility
DNA sequences are available in the NCBI SRA: SAMN02935302 – SAMN02935337. R scripts and code used in the mapping of data: online supporting information.

Supporting Information
Additional Supporting Information may be found in the online version of this article:
Appendix S1. Methods.

Fig. S1 Genome coverage for each synthetic metagenome strain throughout DSN normalization.

Fig. S2 Relative abundance of sequence reads mapped to the 16S rRNA genes within each species, before and after DSN normalization.

Table S1. Pairwise genome distances.

Table S2. Mapping of normalized metagenome sequence reads to reference genomes.

Table S3. Genome enrichment after HAP and DSN normalization.

Table S4. De novo metagenome assembly statistics.

Appendix S2. Perl script for counting mapped reads within Psil fragments.

Appendix S3. BLASTN parsing code for top HSP.

© 2014 The Authors. Molecular Ecology Resources Published by John Wiley & Sons Ltd.