BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 679–686
doi:10.1093/bioinformatics/bts017

Genetics and population analysis

Advance Access publication January 11, 2012

SSuMMo: rapid analysis, comparison and visualization of microbial communities

Alex L. B. Leach*, James P. J. Chong and Kelly R. Redeker*

Department of Biology, University of York, York YO10 5DD, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Next-generation sequencing methods are generating increasingly massive datasets, yet still do not fully capture genetic diversity in the richest environments. To understand such complicated and elusive systems, effective tools are needed to assist with delineating the differences found in and between community datasets.

Results: The Small Subunit Markov Modeler (SSuMMo) was developed to probabilistically assign SSU rRNA gene fragments from any sequence dataset to recognized taxonomic clades, producing consistent, comparable cladograms. Accuracy tests predicted >90% of genera correctly for sequences downloaded from public reference databases. Sequences from a next-generation sequence dataset, sampled from lean, overweight and obese individuals, were analysed to demonstrate parallel visualization of comparable datasets. SSuMMo shows potential as a valuable curatorial tool, as numerous incorrect and outdated taxonomic entries and annotations were identified in public databases.

Availability and implementation: SSuMMo is GPLv3 open source Python software, available at http://code.google.com/p/ssummo/. Taxonomy and HMM databases can be downloaded from http://bioltfws1.york.ac.uk/ssummo/.

Contact: albl500@york.ac.uk

Supplementary information: Supplemental materials are available at *Bioinformatics* Online.

Received on October 12, 2011; revised on December 10, 2011; accepted on January 5, 2012

1 INTRODUCTION

A number of current research foci look to create a better understanding of the complexity of microbial communities and interactions within diverse environments (Rabaey *et al.*, 2007; Raes and Bork, 2008). The analysis of complex microbial communities with high-throughput sequencing (HTS) technologies can generate hundreds of thousands of SSU rRNA reads (Roesch *et al.*, 2007; Sogin *et al.*, 2006; Turnbaugh *et al.*, 2009). SSU rRNA sequences are commonly used to assess community complexity and have been used in such disparate sample regimes as soils (Liu *et al.*, 2008), the human gastrointestinal tract (Ley *et al.*, 2006) and potential biofuel sources (DeAngelis *et al.*, 2011).

As an alternative to primer-targeted studies, whole genome shotgun (WGS) metagenomics has become increasingly popular over the past decade, as it provides additional insight into community function and is purported to reduce sampling bias (Manichanh *et al.*, 2008). Both whole genome and primer-targeted sequencing methods use the same sequencing platforms; technologies producing ever-enlarging datasets (Shendure and Ji, 2008) and suffering similar sequence artefacts, including shorter sequence lengths and greater uncertainty in the prediction of nucleotide bases when compared with older methods (Ledergerber and Dessimoz, 2011).

Regardless of method, it is always desirable to identify those species that most significantly contribute to their environment. Powerful tools to visualize and identify differences or commonalities between datasets, at a number of hierarchical levels, are needed to help understand and model ecosystems and their dynamics in systems biology approaches (Liu *et al.*, 2008; Raes and Bork, 2008). We have developed the Small Subunit Markov Modeler (SSuMMo) in response to the growing computational demands of such large datasets. SSuMMo is based upon a database of profile hidden Markov models (HMMs), trained with the ARB Silva reference database of SSU rRNA sequences (Pruesse *et al.*, 2007). The hierarchy of HMMs (Eddy, 1998) is arranged by EMBL taxonomy and acts as a decision tree to catalogue conserved gene fragments into known species names, one taxonomic rank at a time. This design minimizes the number of pairwise comparisons and bypasses the need to create operational taxonomic units (OTUs), species proxies based on percentage sequence similarity. SSuMMo only groups sequences into acknowledged species names, defined after pure-culture, phenotypic characterizations (Dewhirst *et al.*, 2010; Schloss and Handelsman, 2005).

SSuMMo has been built and optimized for UNIX multicore workstations running Python v2.6+ and is interfaced through a set of command line programs, which can read sequences in over 20 different file formats, as supported by BioPython. SSU rRNA sequences contained within any sequence dataset (genome, HTS gene fragment, etc.) are identified in the first pass of domain-level classifications and retained for further taxonomic classification. Taxonomic assignments can be visualized in real-time, and results automatically saved into a Python object file (Fig. 1A), which is optimized for fast conversion into a number of formats, including phyloxml, html, svg, jpeg, etc. Scripts are provided to calculate abundance and biodiversity information, and fast-track visualization of results using EMBL's IToL web application (Letunic and Bork, 2006), which can paint quantitative and comparative information onto inferred population structures (Fig. 1B). SSuMMo can also save annotated sequences separately, for further downstream analyses, or plot any numeric, tabular data onto the ARB taxonomy (See, Supplementary Method 1.8, Supplementary Fig. S5).

Taxonomic accuracy of SSuMMo was tested by comparing annotated sequences obtained from the NCBI FTP repository

*To whom correspondence should be addressed.

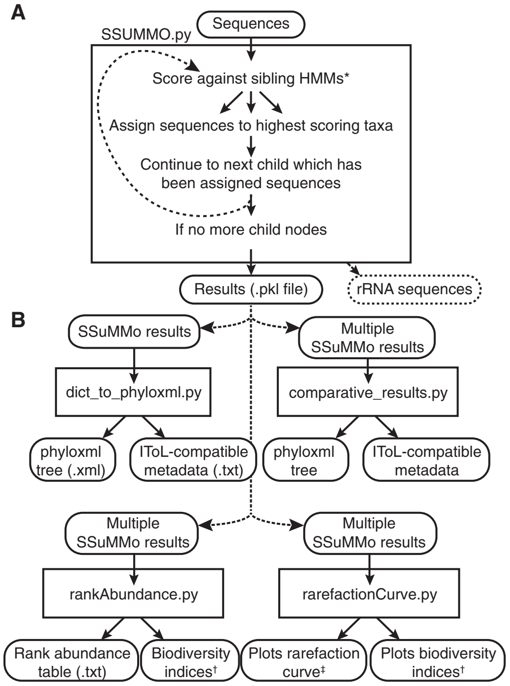

Fig. 1. High level overview of SSuMMo programs. Programs and input/output files are represented with square and rounded boxes, respectively. (**A**) The SSuMMo.py script can accept any sequence type supported by biopython; fasta formatted sequences are expected by default. As the population's taxonomic structure is created, a plain text tree showing quantitative information is printed to screen. Verbose mode also prints hmmsearch tabular results. The main output is a 'pickled file', saved with python's built-in cPickle module, but annotated sequences can be extracted in their original format for further downstream analyses. (**B**) Post-analyses scripts: pickled result files can be used by post-analysis programs to produce various figures and tabular data. Asterisk—sequences are scored against multiple HMMs simultaneously, provided there are spare processor cores. dagger—Simpson and Shannon (H', Hmax) indices are available to choose from. double dagger—rarefaction curves are plotted using python's matplotlib plotting library, and can be saved in raster or vector-based formats.

(ftp://ftp.ncbi.nlm.nih.gov/genomes/TARGET/) and the Human Oral Microbiome Database (Dewhirst *et al.*, 2010) against SSuMMo assignments (Supplementary Figs 1–3, Table 1). Initial tests showed genus prediction accuracy to be >90% (Table 1), prompting development of tools to assist with visualization and comparison of multiple datasets. Functionality was demonstrated with SSU rRNA sequence datasets sampled from lean, overweight and obese individuals (Fig. 3; Supplementary Table S1 and Fig. S4). High-level analyses of pooled results show similar trends to those obtained by thorough analyses performed by Turnbaugh *et al.* (2009), demonstrating SSuMMo's ability to identify trends in dynamic, complex populations.

Further detailed analyses exploring the relative accuracy of assignment in each of nine 'hypervariable' regions in 16S rRNA (V1-9), excised from full and near full-length archaeal test sequences showed targeted sequences as short as 70 nt could identify >70% of genera correctly (Supplementary Figs S2 and 3). Simulations were designed to identify ubiquitously conserved sequence regions suitable for broad-spectrum primers. As HTS methods produce relatively short reads compared with the length of the SSU rRNA gene, we looked to identify those regions in Archaea that coincide with the highest percentage of correct genus predictions (Fig. 2). We note that no single region in SSU rRNA is conserved to an extent as to enable a single primer to cover the entire Archaea domain. Simulated studies could be used to predict those taxa that would be identified with a designed 16S rRNA primer by using the SSuMMo HMM database.

To assist with modelling changes in population structure and diversity within and between datasets, programs were developed to perform rarefaction analyses, calculate biodiversity indices and export stochastic matrices representing taxon probability distributions. Each program can prune resultant taxonomies at any specified rank prior to performing analyses, an alternative to varying cluster sizes by sequence similarity. Results can be exported in tabular form or visualized using Python's matplotlib plotting library (see Supplementary Method 1.6). The provided scripts can apply resampling methods to SSuMMo results,enabling visual comparisons of estimated sampling depth, taxonomic diversity, species evenness and sampling bias within and between datasets. This is performed by 'rarefying', or randomly sampling an equal number of sequences, from result datasets and calculating Shannon and Simpson indices from the observed taxa. Supplementary Figure S4 shows how these methods and metrics can be combined and compared within and between sequence datasets to distinguish high-level features of diversity and community structure.

The ability to combine and visualize species distributions across multiple datasets is a unique feature of SSuMMo, and provides a far speedier alternative to predicting phylogenies, which is prone to human error and can be difficult to reproduce (Peplies *et al.*, 2008). SSuMMo was shown to provide a robust framework for characterization and comparison of population structures, enabling fast access to an array of data dependent metrics. For annotation and inspection, the object-based model provides extensible tools to help compare and edit taxa and sequence annotations between databases.

2 SYSTEM AND METHODS

2.1 Building the SSuMMo database of HMMs

Taxonomy information was parsed from the sequence headers of ARB 'tax' sequence datasets, to create a traversable python object representing sequenced representatives of the tree of life. Due to the size of the uncompressed sequence file (60 GB), an index of sequence locations was created and saved, while simultaneously associating sequence IDs with their relevant species in the python object model (see Supplementary Method 1.1). The ARB Silva (Pruesse *et al.*, 2007) reference alignment of SSU rRNA sequences was made compatible with hmmer, and sequences with gaps or errors were removed. The sequence alignment file was also split by domain, with each produced file processed to remove alignment columns which are gapped in 100% of the domain's sequences. HMMs were trained by all sequences selected from the alignments that are members of each taxonomic group, and were saved in a directory structure created according to ARB's taxonomy (see Supplementary Method 1.2). The model building program (dictify.py) was designed to use a dynamic number of hmmbuild subprocesses that can be used to dramatically accelerate this building stage.

2.2 Associating names with rank

A Python program (link_EMBL_taxonomy.py) was developed to load the latest NCBI taxonomy database and link the taxonomic IDs and ranks to as

Table 1. Species information extracted from fasta sequence headers were compared against SSuMMo taxonomy assignments as a measure of accuracy

Dataset (Rank)	NCBI Archaea[a]		NCBI Bacteria[a]		HOMD extended[b]		HOMD RefSeq[b]	
	Compared (%)	Matched (%)	Compared (%)	Matched (%)	Compared (%)	Matched (%)	Compared (%)	Matched (%)
Phylum	98.6	100	49.8	92.9	43.7	95.1	38.3	97.5
Class	98.6	100	50.1	92.8	58.0	92.2	47.0	95.9
Order	98.6	100	66.1	90.7	72.3	87.4	66.3	93.2
Family	97.2	100	85.2	92.5	74.1	94.5	71.3	96.0
Genus	100.0	97.2	91.5	89.5	78.4	89.1	80.9	85.7
Species	91.7	65.2	94.6	56.8	77.5	44.2	43.1	50.1
# Sequences	144		3 186		34 879		1 646	
Mean Len ± SD	1441.1 ± 36.7		1468.3 ± 47.0		481.7 ± 106.7		1176.3 ± 447.7	

'Compared' shows the percentage of sequence annotations that could be found in the NCBI taxonomy database and propagated back up the tree of life at each rank. 'Matched' shows the percentage of comparable sequences whose rank assignments agreed between SSuMMo and original annotation.

a - ftp://ftp.ncbi.nih.gov/genomes/TARGET/16S_rRNA/.

b - http://www.homd.org/Download - 16S rRNA RefSeq and extended RefSeq databases.

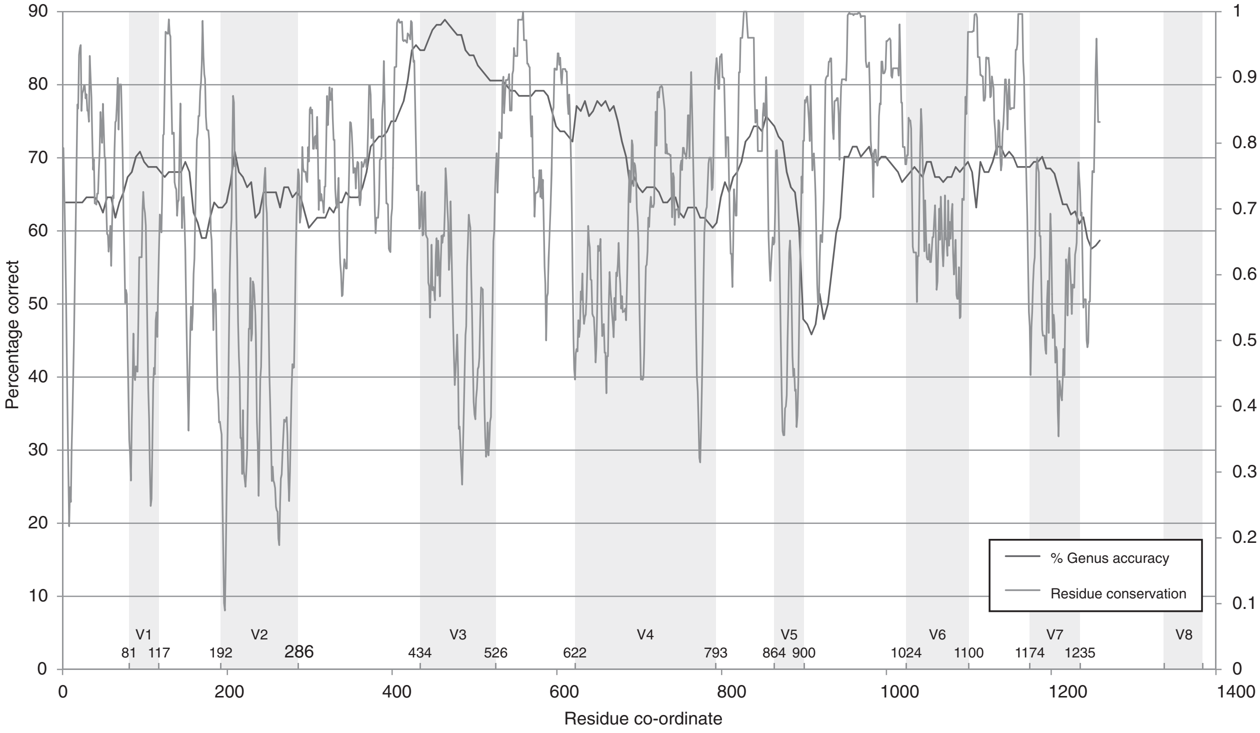

Fig. 2. The percentage of 144 sequences to which SSuMMo correctly assigned genus is plotted against the starting co-ordinate of sequence windows 250 nt in length. Also plotted are C_{10} values, the residue conservation over 10 base windows [Equation (1)], and predicted positions of each hypervariable region for the query sequences.

many ARB taxon names as possible, keeping the associations in a MySQL database. The script automatically downloads and extracts the latest NCBI taxonomy database and loads selected rows (where NameClass = 'scientific name') and columns (tax_ID, name, UniqueName) from the included 'names' table into a local MySQL database. All rows were loaded from the nodes table, but only columns: tax_ID, parent_tax_ID and rank.

New tables for Prokaryotes and Eukaryotes were populated with the ARB taxonomic structure, taking taxon name, parent name, associated NCBI taxonomic ID and rank, wherever the ARB's OTU name/parent name combination uniquely matched. Non-unique name/parent name combinations were inserted into a separate table 'NonUniques' and all IDs recorded. If no match was found for a node, it was given a taxonomic ID of 0 and rank 'unknown' (see Supplementary Method 1.3).

2.3 Assigning novel sequences to taxa

Each query sequence gets scored against profile HMMs in the SSuMMo database one node at a time, choosing the best scoring child of each node as the most probable taxon that the sequence has derived. Starting from the top, each sequence is compared and scored against six profile HMMs: HMMs trained from forward and reverse-transcribed Bacteria, Archaea and Eukaryota sequence alignments. Each query sequence is assigned to the model that returns the highest bit score, according to HMMer v3.0's hmmsearch program. SSuMMo.py continues to recursively traverse the taxon hierarchy, scoring sequences against all HMMs that are direct children of the previous round's assigned taxon. If at any node there are multiple taxa resulting in the same bit-score, SSuMMo will recursively score against all subsequent children from all these equal top-scorers until a unique winner is found. When a clear winner cannot be found, the program will assign the sequence to the last taxon with a unique top-score.

2.4 Accuracy testing

2.4.1 HMM testing Several scoring and model training mechanisms built into hmmbuild were tested to see what effect they had on overall accuracy. HMMs are built using hmmbuild's default options, but HMMs were also built and tested with–wgiven, –max, –nobias and –nonull2 options. wgiven calculates the probability of observing residues in each position directly from the training alignment, whereas by default residue probabilities are calculated with a prior weighting mechanism. max and nobias options affect model sensitivity and acceleration heuristics, and –nonull2 affects the scoring procedure by turning off score corrections based on biased residue compositions.

2.4.2 Sequence length versus Assignment Accuracy The 144 NCBI Archaea sequences were used to test how sequence length affects accuracy of taxon assignment. The full and partial length sequences were shortened at the 3′-end of each sequence by five residues at a time, ensuring that all sequences had identical length, i.e. shorter sequences were removed from the dataset until their sequence lengths were at least the length being analysed. Sequence lengths spanning from 34 to 1509 bases were scored, and NCBI annotations compared with SSuMMo taxon predictions to calculate percentage accuracy according to length (Supplementary Methods 1.3, Supplementary Fig. S1).

2.4.3 SSU rRNA hypervariable region accuracy SSU rRNA hypervariable regions were detected and extracted using Vxtractor (Hartmann *et al.*, 2010). Sequence datasets were synthesized as if primers had been designed to target regions adjacent to each hypervariable region, by extracting sequences of a user-defined length either from the 5′-end or up to the 3′-end of each hypervariable region. Five residues were removed at a time from the opposite end of each sequence window, and the percentage genus accuracy was noted at lengths between 500 and 35 residues (Supplementary Figs S2 and 3).

2.5 Optimizing SSuMMo for speed

A test set of 144 full-length Archaeal rRNA sequences, downloaded from the NCBI ftp servers (ftp://ftp.ncbi.nih.gov/genomes/TARGET/) was used for benchmarking. SSuMMo v0.0.1 worked on a one-to-one basis, parsing one sequence at a time and scoring that sequence against a single profile HMM using hmmsearch.

SSuMMo v0.0.2 worked on a many-to-one basis, perceived as such because all sequences are scored against a single model at a time, again using HMMer v3.0's hmmsearch.

SSuMMo v0.0.3 was built with a many-to-many sequence-model comparison in mind, by using HMMer v3.0's hmmscan. In order to use hmmscan, the SSuMMo database had to be modified to include 'pressed' collections of HMMs. In order to facilitate this database update, dictify.py was extended to optionally use hmmpress on all HMMs at a given node. Upon updating the database, SSuMMo v0.0.3 was updated to use hmmscan, scoring all sequences at a node to that node's pressed collection of HMMs in a single program call. The aforementioned set of 144 sequences were used to test all versions of SSuMMo and times taken for analysis compared (data not shown). SSuMMo v0.0.2 was found to be the quickest implementation and was selected for further development to utilize multiple processors.

2.6 Comparative metagenomics

A Python program (comparative_results.py) was written to combine SSuMMo results files and show community differences in terms of diversity, ubiquity and abundance. Phyloxml formatted trees can be exported and programmatically uploaded to ITOL (Letunic and Bork, 2006), with delimited data files showing population structure and community differences, which can be co-represented on cladograms as multi-value bar graphs. (e.g. http://itol.embl.de/external.cgi?tree=226561982157513085564600). Multiple sequence files can be grouped and the ubiquity of species across each group exported as tabular form or ITOL representation as heatmaps. The user also has the option of programmatically downloading the tree again in any of the formats ITOL allows to be exported (pdf, jpeg, etc.; see Supplementary Methods).

3 RESULTS AND DISCUSSION

3.1 Assignment accuracy

Initial accuracy tests were performed with 144 full and near-full length Archaeal 16S rRNA sequences (all >1257 bp) obtained from ftp://ftp.ncbi.nih.gov/genomes/TARGET/. Up to 99% (142) were assigned to the correct genus and 100% of sequences are correctly assigned to higher ranks, according to their original NCBI annotation (Table 1). No difference in accuracy was noted between the different model training methods, when using the Archaea test dataset. However, we found that hmmbuild's default settings made HMMs giving the best accuracy when using the NCBI Bacteria dataset of full length 16S rRNA sequences.

The impact that sequence length had on SSuMMo's assignment accuracy was investigated with the same test dataset, by trimming residues from the 3′-end of aligned sequences, before analysing with SSuMMo, and tallying the scores (Supplementary Fig. S1). Interestingly, genus assignment accuracy increased to the maximum of 99% (142) only after trimming the last 85 residues from the 3′-end of the test sequences. At lengths between 1119 and 1364 residues, SSuMMo assigned sequences with a genus accuracy of 98%, below which accuracy declined in a non-linear fashion (Supplementary Fig. S1). SSuMMo genus assignment accuracy was <95, 90, 80 and 70% for sequence lengths of 1059, 959, 554 and 387 ± 2 residues, respectively.

Further tests were performed on SSU rRNA hypervariable regions, as detected by V-Xtractor (Hartmann *et al.*, 2010) (Supplementary Figs S2 and 3), by extracting sequences extending 500 residues to or from locations either side of each hypervariable region. SSuMMo was iteratively run on sequences after shortening by five residues at a time, and percentage accuracies recorded. Our results show that the V4 region most accurately assigned genera throughout the domain, with accuracies remaining $\geq$ 75% for sequence lengths of just 67 ± 2 residues (Supplementary Figs S2 and 3; raw data not shown). The V9 region consistently performed worst, which is likely explained by a lack of training data, as many of the Archaea sequences in the ARB database do not cover this region, which spans alignment columns 1310–1340 [according to alignments against RNAMMER HMMs (Lagesen *et al.*, 2007)].

Some of the lowest accuracies for assignments within the Archaea domain occurred with regions at the 3′-end of the full sequences (Fig. 3, Supplementary Figs S2 and 3). This can be explained by the increased likelihood of errors appearing at the tail of sequence reads (Flicek and Birney, 2009) and by the fact that many training sequences were not full length. Out of 511 814 training sequences

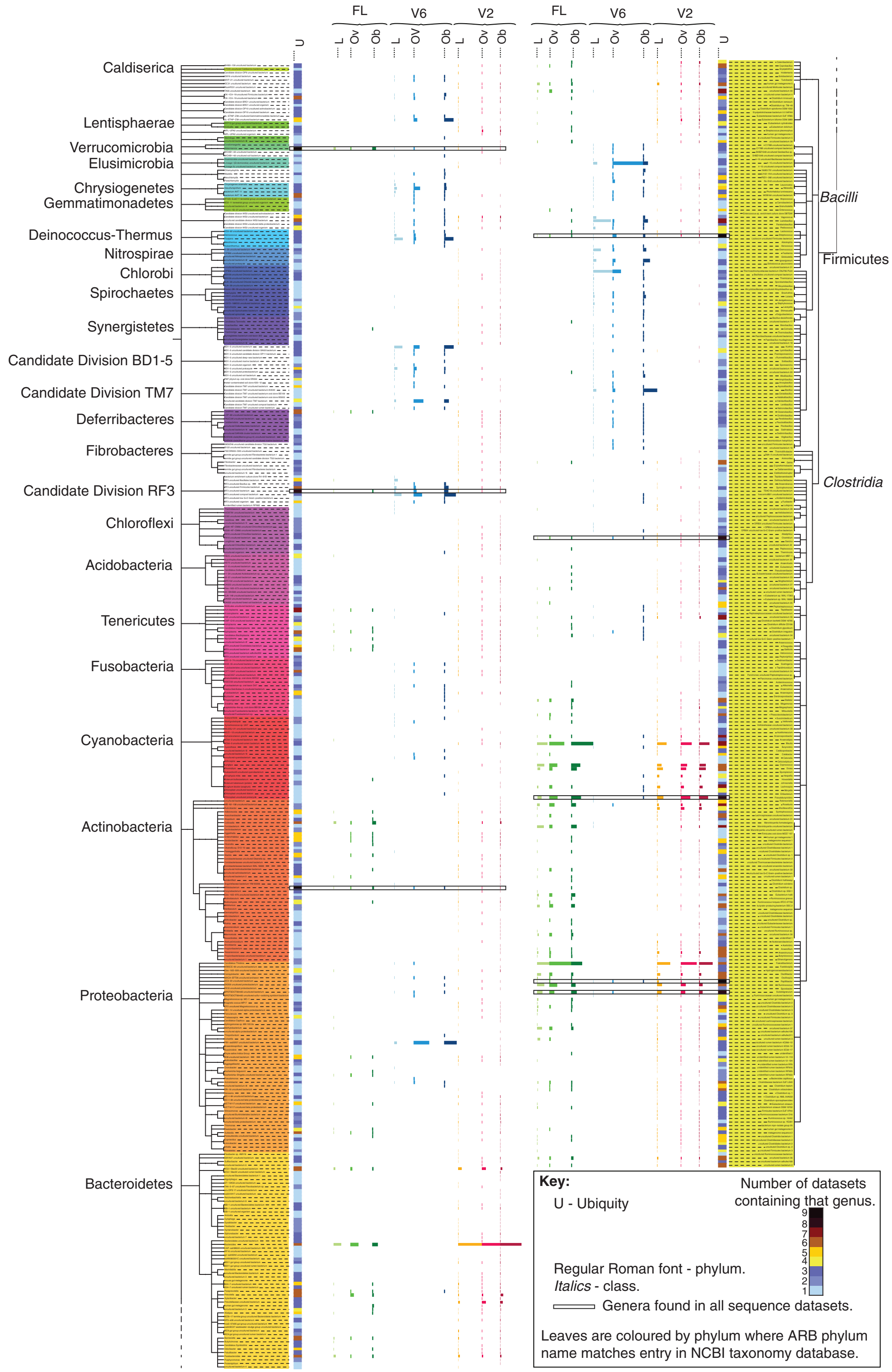

Fig. 3. Distribution of taxa up to genus specificity, present in the guts of 154 lean, overweight and obese individuals, pooled by sequencing method and body mass index (BMI) category. Graphs are shown grouped in order of increasing number of sequences generated per PCR-method, and represent the relative abundances of each taxon that were identified in that sequence pool. FL—Full Length sequences, V6 & V2—sequences generated from V6 & V2 region specific primers. L, Ov, Ob—Lean, Overweight and Obese BMI categories.

housed in ARB v104 database, 9667 sequences are < 1200 residues in length, the majority of which are members of the Archaea domain (9621), representing > 45% of the 20 994 Archaea sequences in the ARB v104 database.

At sequence lengths of 400 nt, a common read length generated by pyrosequencing technologies (Droege and Hill, 2008), SSuMMo was shown to accurately predict the genus of > 70% of archaeal sequences targeted at either end of regions V1-6 (Supplementary Figs S2 and 3). Methodologies producing even shorter reads would benefit from well-designed primers, as accuracies as high as 80.5% are achieved with sequences 250 bp in length, if starting from the 5′-end of the V4 region (Supplementary Fig. S2).

SSuMMo accuracy was tested for consistency in the Bacteria domain using sequences obtained from NCBI and the Human Oral Microbiome Database (HOMD)(Dewhirst *et al.*, 2010). SSuMMo correctly assigned >86% of reference sequences to genera described in sequence annotations (Table 1), and > 92% of bacterial family predictions matched their annotation across all datasets, up to 96% accuracy for the HOMD RefSeq database. The lowest accuracies were recorded for species level assignments. A random sample of 100 mis-assignments indicated ~40% were being assigned to uncultured species, with about half of those being assigned to the correct genus (Supplementary Table S3). The mis-assignments could be due to a number of factors, including subtle differences in naming conventions between databases (we estimate ~25% of mis-assigned sequences), differences in the number of training sequences, and the taxonomy structure which underlies our method (ARB has multiple unclassified branches at many different nodes).

Matching taxa names between databases posed a problem as database entries are often misspelled (e.g. in ARB: Brumimimicrobium instead of Brumimicrobium etc.), mismatched (e.g. exchangeable, non-alphanumeric characters), or non-unique (e.g. Acidobacterium is both a phylum and a class). These issues do not affect SSuMMo's ability to assign sequences to most probable taxa, but negatively affect the inferred number of comparable sequences in the accuracy tests (Table 1).

3.2 Software comparisons

SSuMMo processing times were compared with those of BLASTN and MEGABLAST (v2.2.21) using an array of datasets (see Suppl. Materials). SSuMMo took 4 hours, 7 mins to process 291 993 V2-targeted sequence reads and 6 h 32 min to process 3186 near-full-length sequences (Supplemantary Table S2). When compared against the default BLAST configurations (1 CPU core), SSuMMo is fastest, but after changing BLAST settings to use 11 of 12 CPU-cores, as SSuMMo did by default, MEGABLAST was fastest with datasets up to several thousand sequences, but slower than SSuMMo with the largest tested dataset (Supplementary Table S2).

SSuMMo's accuracy (Table 1) outperforms tools used to annotate WGS metagenomic datasets (Brady and Salzberg, 2009), as expected, given that SSU rRNA is the most highly sequenced gene, by far. The RDP classifier, designed specifically to annotate SSU rRNA sequences reports comparable taxon prediction accuracies (Wang *et al.*, 2007).

$$C_n^{n+i} = \frac{1}{i}\sum_{n=1}^{n+i} 4 \cdot Var \cdot (p_A, p_C, p_G, p_T) \qquad (1)$$

3.3 Sequence windows and primer design

Prokaryotes contain nine hypervariable regions in their 16S rRNA gene, which are interspersed with relatively conserved regions that are more suitable for designing broad-spectrum PCR primers. SSuMMo was tested to see if the extra variation in these regions affected genus predictions, by excising a 250 base 'window' within each archaeal sequence and shifting it 5 nt at a time (Fig. 2). In this scenario, the highest accuracy recorded within this set of 144 sequences was 89% and the lowest was 48%. The nucleotide conservation in Archaea sequence alignments was calculated and averaged over 16 base windows along the whole SSU rRNA gene (Equation 1; $I = 16$). This returns a value between 0 (no conservation) and 1 (perfectly conserved region) for any group of aligned sequences. The start position of the most accurately assigned 250 base window was identified in the middle of the V3 hypervariable region, where SSU rRNA gene is low (Fig. 2), making this an unsuitable location for targeted primer design. A more effective primer selection might focus upon RNAMMER alignment positions 535–551, between regions V3 and V4 as it is highly conserved ($C_{16} = 0.992$) (5′-CAGC[-c][AC]GCCGCGGUAA-3′). There are three 250 base long sequence windows, starting from local alignment positions 562, 567 and 572 and extending downstream, which show accuracies of 79%; the highest accuracy for any region starting from a ubiquitously conserved region of sufficient length for primer design. However, if targeting the reverse strand from this location, typical sequence lengths would extend beyond the V3 region into positions that are relatively worse at resolving taxa accurately.

3.4 Identifying community differences

Targeted HTS has become a popular method for characterizing microbial populations present in the human microbiome (Adlerberth and Wold, 2009; Dewhirst *et al.*, 2010; Turnbaugh *et al.*, 2009). Sequence reads sampled from 154 lean, overweight and obese twins and their mothers (Turnbaugh *et al.*, 2009) were used to test SSuMMo's applicability to such studies. The sequences were used to compare global trends in the data, according to BMI category and sequencing method. SSuMMo analyses of V2 regions and full-length sequences concur with the original observations (Turnbaugh *et al.*, 2009): that obese subject samples have significantly fewer Bacteroidetes, more Actinobacteria and no significant difference in Firmicutes relative to lean individuals (Supplementary Table S1). Similar trends were observed across the dataset at finer resolutions, with no single genus dominating any subset of the data (Fig. 3).

The same SSuMMo results were used to find ubiquitously conserved taxa across Turnbaugh *et al.*'s data (2009). Conserved taxa are visualized as 'color strips' on the IToL web application, so as to quickly and easily identify conserved taxa (Fig. 3). Across all sampling methods and BMI categories only eight genera: *Akkermansia, Bifidobacterium, Streptococcus, Clostridium, Pseudobutyrivibrio, Papillibacter, Subdoligranulum* and uncultured members of Candidate Division RF3 were found in all result sets (Fig. 3). Some of these genera have been reported as beneficial to health when found in human intestinal tracts [e.g. *Bifidobacterium* (Hao *et al.*, 2011), *Akkermansia* (Derrien *et al.*, 2007), etc.]. However, full genome information will be needed to elucidate if each provides unique metabolic capabilities that justify their ubiquitous nature.

Primers are known to anneal favourably with certain taxa over others (Chakravorty *et al.*, 2007), leading to a sampling bias dependent on the DNA sampling method. This effect is apparent in Turnbaugh *et al.*'s results, where presence and absence information show V2 and V6-specific primers to have more influence on observed population structure than host BMI category (Supplementary Table S1; Fig. 3). Although V6 taxon assignments appear anomalous compared with assignments based on V2 fragments and full-length sequences, V6 results show high resolution in members otherwise missed. This is demonstrated by the fact that the V6 sequence data identified so few Bacteroidetes sequences, even though it is the second most abundant phylum in all other sequence sets (Supplementary Table S1). Similar evidence at the class level is observed, as many members of the class *Bacilli* are ubiquitously present in all V6 sequence sets in high proportions, yet are not present in the other sequence datasets at all. Conversely, many members of the class *Clostridia*, also in the phylum *Firmicute*s, are observed in high proportions with full-length and V2 sequences, but are not identified at all with V6 reads.

3.5 Biological diversity

As with all other SSU rRNA identifying software, SSuMMo does not account for multiple rRNA operon copy numbers per genome, which vary between 1 and 15 copies per organism, where this information is available (Supplementary Fig. S5) (Klappenbach *et al.*, 2000). There is also variation in chromosomal copy number between organisms, which may vary with proliferation state (Pecoraro *et al.*, 2011). These factors mean that quantifying 16S rRNA genes in environmental samples does not indicate the number of individual cells in a sample, but only the number of rRNA gene copies sequenced. Together, these could contribute a 2–3 order of magnitude error in organism estimates.

Using rank abundance scores and information gained on the species distributions, several biodiversity indices can be calculated, at distance thresholds defined by taxonomic rank rather than percentage sequence similarity, which is commonly used when defining OTUs (e.g. (Schloss and Handelsman, 2005)). Shannon and Simpson biodiversity indices, biological diversity measures incorporating evenness and richness, respectively (Magurran, 2009), were calculated for each BMI category based on species-level taxa assignments (Supplementary Table S1). These results were used to investigate whether notable changes in biodiversity could be identified when sequences were grouped at species rank. No consistent changes were observed across all three BMI categories and sequence targets, as pooled samples obfuscate more subtle differences which might be observed between individuals. For example, gut populations were shown to be more similar between family members, so characterizing populations from lean and obese members of the same family (rather than all families pooled together) would be a fairer method of delineating differences between BMI categories. Furthermore, variation in the number of defined species per genus across the tree of life will cause differences in primer specificity to drastically affect Shannon and Simpson index calculations, which are affected by the number of observed species groups.

In order to correct for differences between sequence sample sizes, we applied rarefaction analyses to each member dataset, to select random samples from each member dataset. By plotting calculated Shannon and Jackknife indices from random subsamples of Turnbaugh *et al.*'s data, we observe trends in the V2 and full-length sequence datasets that appear to follow with the size of each set of sequences. As mentioned above, these trends are likely affected by the number of individuals sampled and pooled into a combined sequence dataset, as with more sampled individuals, more singleton taxa are introduced. The V6 dataset is unique in that there are fewer sequences in total sampled from lean individuals, yet more genera are observed (Supplementary Fig. S4). This corresponds with a slightly higher species evenness, or Shannon H' value (Supplementary Table S1, Supplementary Fig. S4D), and a noticeably higher $H_{\max}$ value, suggesting that those taxa targeted by V6 primers (Fig. 3) are more evenly distributed in Lean individuals than in their counterparts with higher BMI ratios.

3.6 Repository annotation effects

SSuMMo relies upon public repository data to generate its model libraries and taxonomy information, and is therefore sensitive to inaccurate or outdated sequence annotations present in public repositories (Siezen and Van Hijum, 2010). Inaccuracies and inconsistencies between databases reduce inferred assignment accuracies, but these difficulties are faced by all software which rely on pre-existing data to classify new sequences. Through working with SSuMMo and the annotated test datasets, various inconsistencies were observed between sequence annotations and species names found within the ARB database. Often, annotated sequence names could not be found in the ARB database, with further investigations showing the most likely causes to be human error, asynchronous name-changes or taxa deliberately introduced into one database and not the other. The percentage of uncultured species described in the ARB and NCBI databases is sizeable, with 11 126 and 15 200 taxa names starting 'uncultured', respectively. Many taxa have numerous versions of uncultured species too. For example, the family Methanobacteriaceae contains four variations on 'uncultured' in the ARB database, including 'uncultured', 'uncultured archaeon', 'uncultured Methanobacteriales archaeon' and 'uncultured Methanobacteriaceae archaeon'. The NCBI taxonomy contains all of these names just once, but none of them appear as children to Methanobacteriaceae.

Prior to isolating a culture, formal species names cannot be accurately assigned due to an inability to fully characterize an organism's phenotype (Dewhirst *et al.*, 2010). This suggests that these uncultured species have been predefined based on (dis)similarity of SSU rRNA sequences alone. As more extensive information is determined about species whose sequences are defined as uncultured, eventually leading to the definition of new species, it will be a challenge to maintain and update public databases while assigning 'uncultured' sequences to their appropriate names.

Many of these uncultured species are direct children of a family name, e.g. the family Halobacteriaceae is parent to the species 'uncultured archaeon', skipping the genus level assignment and therefore bypassing the rank that SSU rRNA can confidently be assigned. These curatorial discrepancies cause difficulties when trying to assess the accuracy of SSuMMo (or any similar methods) using name-based matching between taxa.

3.7 SSuMMo for database curation

SSuMMo shows extremely high accuracies at ranks higher than genus. We suggest that current sequence and taxonomy databases may benefit from features of SSuMMo that assist with fast identification of outdated and erroneous entries. This would benefit individuals and database administrators to achieve consistency when describing sequence taxonomies and phylogenetic mappings. Consistency checks could be incorporated both pre- and post-submission of SSU rRNA sequences into public repositories. The read sizes produced by next-generation sequencing methods enabled datasets containing hundreds of thousands of SSU rRNA sequence reads to be allocated to taxa in several hours (Supplementary Table S2). Running SSuMMo on a raw dataset could assign sequences to probable taxa quickly and effectively, and would also give extra assurance to annotations made with any other method.

Sequences already annotated in public repositories would also benefit from the assurance of a correct SSuMMo allocation. Not only are scripts provided to download and update the latest NCBI taxonomy database and load a minimized version into MySQL, but annotations can be compared with real taxa with their corresponding rank and NCBI taxonomic ID. As the EMBL SSU rRNA database continues to be updated and enlarged, the reference collection of SSU rRNA sequences will continue to grow, and so will the ARB Silva database of aligned SSU rRNA sequences. ARB v106 currently has 1.9 million 16S rRNA sequences and the reference database over 500 000 high-quality, aligned sequences allocated to 134 956 nodes across all three domains of life. As these databases continue to grow exponentially, SSuMMo's database will not, yet it will still be updated to incorporate the latest sequence data released with EMBL, and subsequently ARB. Instead of growing (and performance decreasing) with the release of new reference sequences, SSuMMo will only continue to grow with newly defined taxa, which will only become more informative and accurate in their assignments.

ACKNOWLEDGEMENTS

We thank the James Burgess Scholarship and Holbeck Trust for support (to A.L.B.L.). We would also like to thank Stuart Priest and the Technology Facility for use of their computing facilities and Martin Hartmann (University of British Columbia) for sharing custom, in-house scripts.

Funding: This work was funded by a James Burgess Scholarship, supported by The Holbeck Trust.

Conflict of Interest: none declared.

REFERENCES

Adlerberth,I. and Wold,A.E. (2009) Establishment of the gut microbiota in Western infants. *Acta Pædiatrica*, **98**, 229–238.

Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Meth.*, **6**, 673–676.

Chakravorty,S. *et al.* (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, **69**, 330–339.

DeAngelis,K.M. *et al.* (2011) Characterization of trapped lignin-degrading microbes in tropical forest soil. *PLoS ONE*, **6**, e19306.

Derrien,M. *et al.* (2007) The Mucin-degrader Akkermansia muciniphila is an abundant member of the human intestinal tract. *Appl. Environ. Microbiol.*, **74**, 01226–01207.

Dewhirst,F.E. *et al.* (2010) The human oral microbiome. *J. Bacteriol.*, **192**, 5002–5017.

DroegeM, and Hill,B. (2008) The Genome Sequencer FLX(TM) System–Longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotech.*, **136**, 3–10.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Flicek,P. and Birney,E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Meth.*, **6**, S6–S12.

Hao,Y. *et al.* (2011) Complete genome sequence of Bifidobacterium longum subsp. longum BBMN68, a new strain from a healthy Chinese centenarian, *J. Bacteriol.*, **193**, 787–788.

Hartmann,M. *et al.* (2010) V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16 S/18 S) ribosomal RNA gene sequences. *J. Microbiol. Meth.*, **83**, 250–253.

Klappenbach,J.A. *et al.* (2000) rRNA Operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.*, **66**, 1328–1333.

Lagesen,K. *et al.* (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

Ledergerber,C. and Dessimoz,C. (2011) Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* [Epub ahead of print, doi:10.1093/bib/bbq077, January 18, 2011].

Letunic,I. and Bork,P. (2006) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Ley,R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature*, **444**, 1022–1023.

Liu,Z. *et al.* (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.*, **36**, e120.

Magurran,A.E. (2009) *Measuring Biological Diversity*. Blackwell Publishing, Oxford.

Manichanh,C. *et al.* (2008) A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. *Nucleic Acids Res.*, **36**, 5180–5188.

Pecoraro,V. *et al.* (2011) Quantification of Ploidy in Proteobacteria Revealed the Existence of Monoploid, (Mero-)Oligoploid and Polyploid Species. *PLoS ONE*, **6**, e16392.

Peplies,J. *et al.* (2008) A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. *Syst. Appl. Microbiol.*, **31**, 251–257.

Pruesse,E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.

Rabaey,K. *et al.* (2007) Microbial ecology meets electrochemistry: electricity-driven and driving communities. *ISME J.*, **1**, 9–18.

Raes,J. and Bork,P. (2008) Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Micro.*, **6**, 693–699.

Roesch,L.F.W. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.*, **1**, 283–290.

Schloss,P.D. and Handelsman,J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, **71**, 1501–1506.

Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotech.*, **26**, 1135–1145.

Siezen,R.J. (2010) Genome (re-)annotation and open-source annotation pipelines. *Microbial Biotechnology*, **3**, 362–369.

Sogin,M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.