

ANALYSIS OF SOIL SAMPLE METAGENOME AS AN ADDITIONAL TOOL FOR FORENSICS

Aaspõllu A¹, Lilje L¹, Simm J¹, Kägo E¹, Sipp Kulli S², Moora M³, Zobel M³, Metsis M^{1,2}

¹Tallinn University of Technology, Tallinn, Estonia; ²BiotaP LLC, Tallinn, Estonia; ³University of Tartu, Tartu, Estonia

anu.aaspollu@gmail.com



19th IAFS World Meeting
Funchal - Madeira- Portugal
12-17 September 2011

INTRODUCTION

Soil analysis is a valuable resource in legal investigation potentially connecting victim or suspect to a particular crime scene because soil traces are easily transferred to persons and/or objects. Classical forensic soil analysis involves examination of its physical characteristics and chemical composition, such as soil type, colour, particle size, pH, elemental, mineral and organic content (Marumo et al 1996). However, the limited variability of these parameters does not always allow adequate discrimination between soil samples. Therefore, microbiological approaches have been proposed as technique that could fulfil the gap and provide complementary information. To date terminal restriction fragment length polymorphism (T-RFLP) analysis has been the most popular technique for microbial-based soil identification and differentiation (Lenz and Foran 2010; Macdonald et al 2011).

The aim of current study was evaluation of variability of bacterial and eukaryotic (including fungal) communities in different soil samples on metagenome level using the second generation sequencing technique for development an additional tool for linking persons to the crime scenes.

RESULTS

Table 1. Mean numbers of sequences obtained per samples using different systems (16S, 18S and AMF) and mean numbers of matched sequences found in NCBI database.

Sample code	16S		18S		AMF	
	Mean # of sequences	Mean of matched sequences	Mean # of sequences	Mean of matched sequences	Mean # of sequences	Mean of matched sequences
INT1	1387	0.71	2659	0.69	1503	0.80
KH1	1121	0.70	215	0.60	1028	0.48
KO	1005	0.73	2036	0.69	420	0.42
KTM2	1149	0.56	4421	0.88	2270	0.84
KU2	4685	0.78	1857	0.74	573	0.40
Mahe1	1081	0.69	2283	0.69	1208	0.61
PU	1261	0.72	1948	0.67	335	0.45
Ray	1017	0.73	2671	0.64	435	0.66
TR	633	0.74	1817	0.69	460	0.40
JSI1	1504	0.73	2728	0.80	207	0.18
VI	656	0.67	2113	0.66	546	0.66

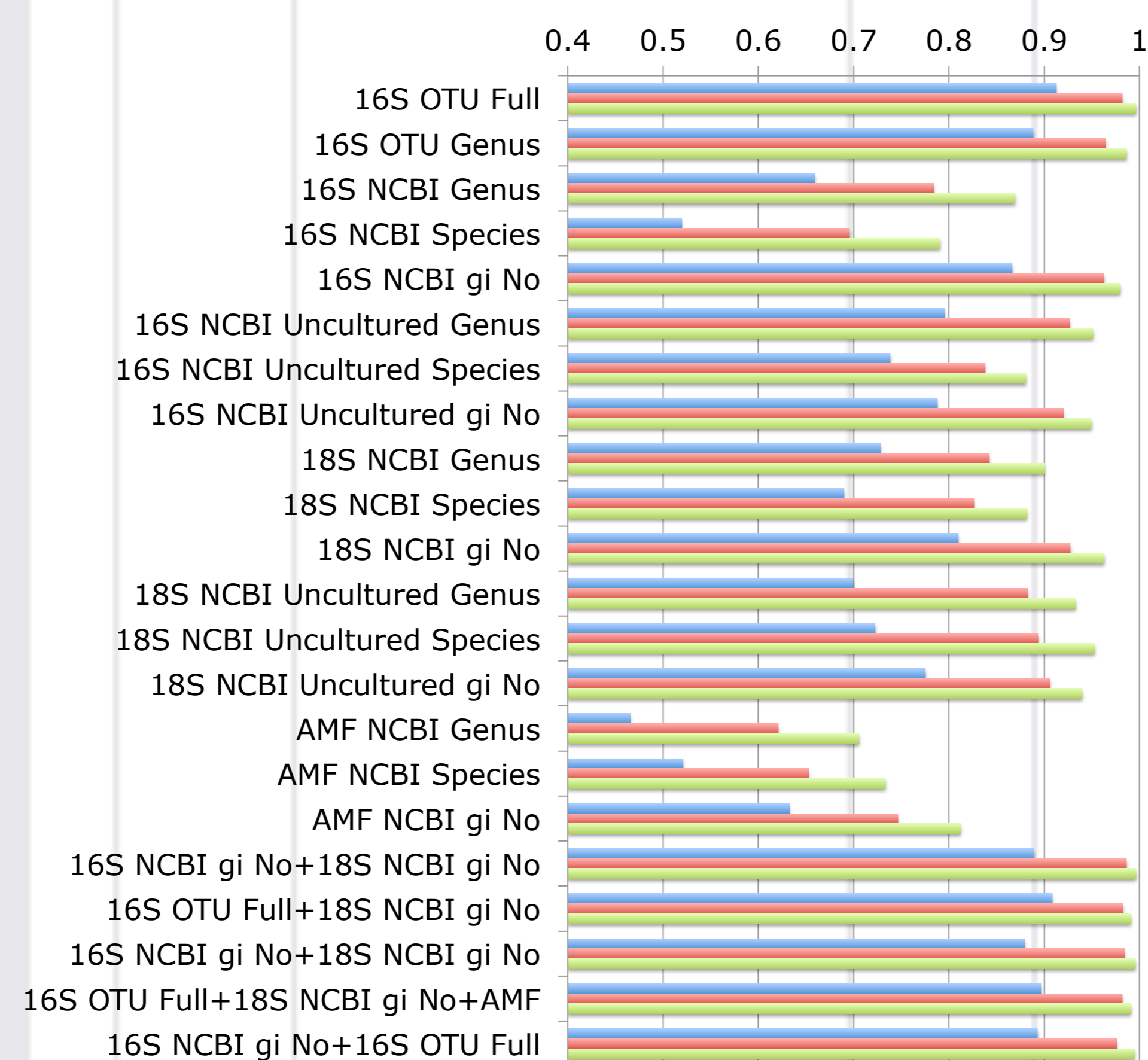


Figure 1. Accuracy of classification for 9 sampling areas for finding true area (average of 20 trials).

- **Top 1** score (accuracy) shows the probability that the true area is the predicted area.
- **Top 2** score shows the probability that the true area is in the first 2 predicted areas.
- **Top 3** score shows the probability that the true area is in the first 3 predicted areas.

Approach used	Top 1	Top 2	Top 3
16S OTU Full	0.9124	0.9818	0.9961
16S NCBI gi No+18S NCBI gi No	0.8890	0.9864	0.9961
16S OTU Full+18S NCBI gi No	0.9084	0.9825	0.9909
16S NCBI gi No+18S NCBI gi No+AMF NCBI gi No	0.8792	0.9844	0.9955
16S OTU Full+18S NCBI gi No+AMF NCBI gi No	0.8961	0.9818	0.9909
16S NCBI gi No+16S OTU Full	0.8929	0.9760	0.9948

Table 2. Best performance or statistically non-different to best performance.

Sample code	INT1	KH1	KO	KTM2	KU2	Mahe1	PU	Ray	TR
INT1	8	0	0	0	0	0	0	1	0
KH1	0	9	0	0	0	0	0	0	0
KO	0	0	6	0	1	1	0	1	0
KTM2	0	0	0	8	1	0	0	0	0
KU2	0	0	0	0	9	0	0	0	0
Mahe1	0	0	2	0	1	6	0	0	0
PU	0	0	0	0	0	0	9	0	0
Ray	0	0	0	0	0	0	0	9	0
TR	0	0	0	0	0	0	0	0	8
JSI1	0	0	8	0	1	0	0	0	0
VI	1	1	0	0	0	0	2	4	1

Table 3. Confusion table based on 16S OTU full data of known samples for cross validation of statistical model. Each row shows the predicted areas for 9 samples from particular area.

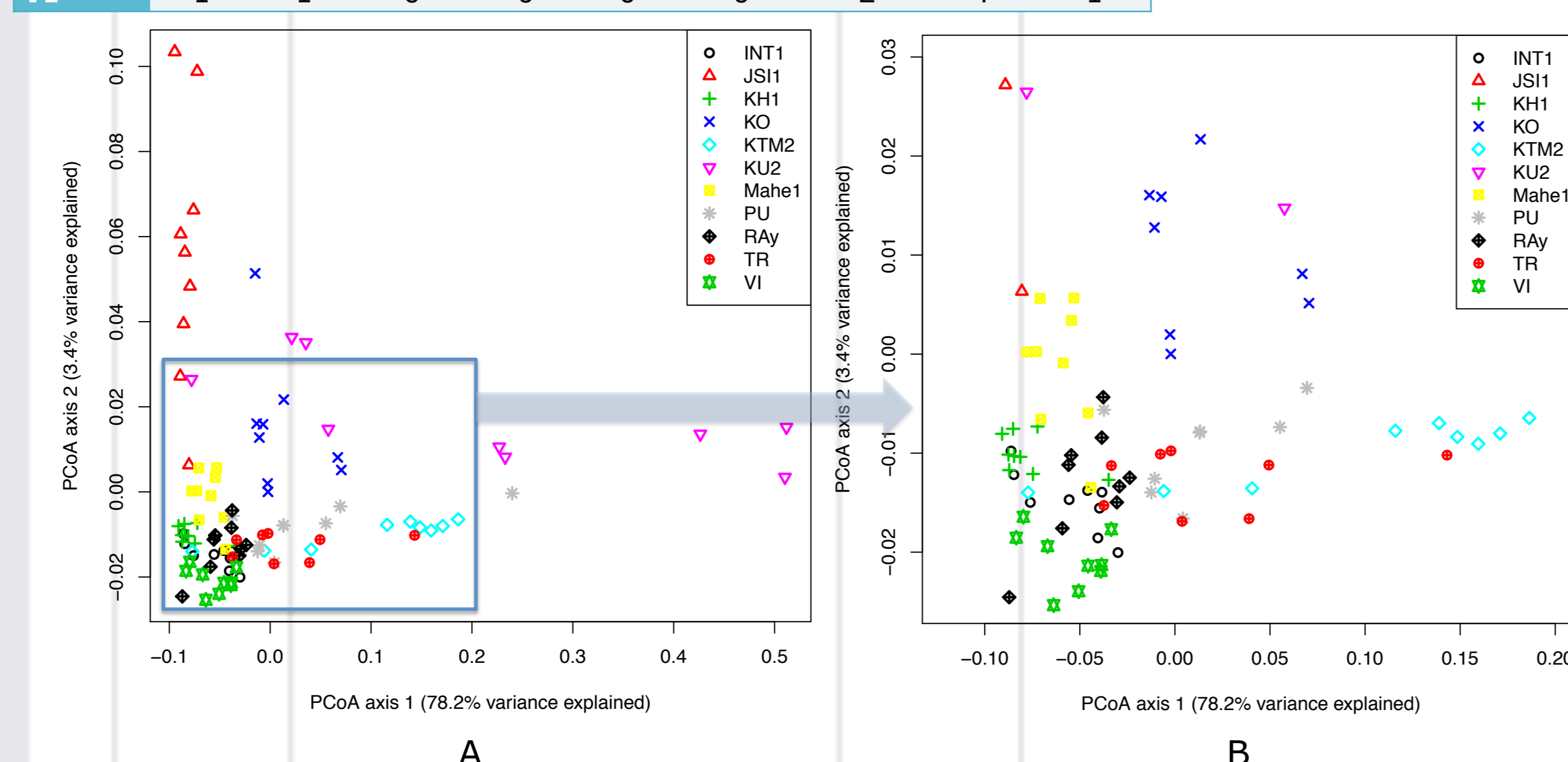


Figure 2. PCoA (Principal Coordinate Analysis) for 16S OTU Full based distribution of sampling areas/samples. A. Original plot; B. Zoomed plot area.

MATERIALS AND METHODS

Soil samples were collected from different environments with different flora. 9 soil samples per one sampling area (30m x 30m) spaced by 15m were subjected for analysis.

Table 4. Sampling areas.

Sample code	Geographical location	Sample type
INT1		Intensively cultivated field
KH1	Virtsu	Cocksfoot field
KO	Koeru	Eutrophic boreo-nemoral forest
KTM2	Kiviõli	Gangue hill
KU2	Kurtna	Successive forest resulting from human intervention
Mahe1		Organic food field
PU	Laelatu	Wooded meadow
Ray	Rasina	Field of rapeseed
TR	Tartu Riia street	Park
JSI1	Järvselja	Spruce plantation
VI	Virtsu	Fallow

Samples used for validation of statistical model are in bold (so called unknown samples).

DNA was extracted using PowerSoil® DNA Isolation Kit (MO BIO Laboratories, USA) according to manufacturer's recommendations. For bacterial community analysis the 16S rRNA gene V2-V3 hypervariable region was amplified using primers 8F and 357R, for fungal community sequencing the 18S rRNA gene V2-V3 region was amplified using primers 4EF and 518R, and for arbuscular mycorrhizal fungi (AMF) NS31-AML2 SSU rRNA region using primers AML1 and AML2 (Table 2).

Table 5. Primer sequences

Primer	Sequence 5'-3'	Reference
8F	TTGGCAGTCTCAGnnnnnnnn AGTTTGATCCTGGCTCAG	Armougom and Raoult 2009
357R	GTCTCCGACTCAGnnnnnnnn CTGCTCCCTYCCGTA	Schabereiter-Gurtner et al 2001
4EF	GTCTCCGACTCAGnnnnnnnn GGAAGGGRTGATTATTAG	Lee et al 2008
518R	TTGGCAGTCTCAGnnnnnnnn ATTACCGCGGCTGCTGG	
AML1	TTGGCAGTCTCAGnnnnnnnn ATCAACTTTCGATGGTAGGATAGA	
AML2	GTCTCCGACTCAGnnnnnnnn GAACCCAAACACTTTGGTTCC	

Underlined sequence denote 454 specific sequencing primer parts B and A respectively, "n" decode 8-bp barcode sequence and particular gene specific sequence is in bold.

Parallel DNA sequencing was performed on Roche/454 platform using 454 GS FLX Titanium System.

For data analysis two computational models based on statistical analysis were developed. The first model employs operational taxonomic unit (OTU) definition by clustering of sequences based on mothur software (Schloss et al 2009) and the second one uses NCBI BLASTN. For both approaches 9 samples per one sampling area were used for building up a statistical model of the community. Modelling accuracy was estimated throughout cross-validation, excluding one sample per sampling area and running analysis 9 times. We created different datasets using:

- All sequences (OTU Full, NCBI gi number based)
- Limited sequences (genus defined, species defined, uncultured organisms).

Unknown samples from two additional sampling areas (forest JSI1 and field VI) were analysed to test the performance of our system.

DISCUSSION AND CONCLUSIONS

To overcome limitations related to T-RFLP analysis, concerning different sizing results obtainable from different instruments and analysis conditions, we implemented direct sequencing approach to get more precise taxonomic information on species/genus level. Thus, this parallel sequencing approach enables to obtain reliable and comparable information about existing microbial communities that could be shared very easily with other (forensic) laboratories. As the approach uses direct sequencing, it meets also quality criteria for accreditation, which is an important question for forensic purposes. Microbial community analysis of soil samples in the current study revealed that samples from different environment were significantly distinguishable, although more samples have to be analysed including samples taken from the same locations but at different times/seasons as well as from wider range of soil-types. Nevertheless, our proposed statistical modelling approach using logistic regression showed good performance for used dataset.

Two unknown samples, which were used for testing performance of our system, performed well: the first unknown sampling area (Järvselja spruce plantation) found some similarities to Koeru eutrophic boreo-nemoral forest and Kurtna successive forest resulting from human intervention, the only forest areas used for building up statistical model. The second unknown sampling area (Virtsu fallow) did not find similarities to Koeru eutrophic boreo-nemoral forest, Kiviõli gangue hill, Kurtna successive forest resulting from human intervention and organic food field.

The most informative statistical approach for finding true sampling area was based on OTU clustering using all sequences available (16S bacterial communities).

Thus, the obtained results are clearly promising enabling exclude or include samples to particular site; however, foresee a need for more elaborative studies for statistical evaluation and criteria formation to be able to implement this approach into the routine forensic practice.

REFERENCES

Armougom F., Didier R. (2009) Exploring Microbial Diversity Using 16S rRNA High-Throughput Methods. *Journal of Computer Science and Systems Biology*. 2(1): 074-092.

Lee J, Lee S, Young JP. (2008) Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. *FEMS Microbiol Ecol*. 65 (2): 339-49.

Lenz EJ, Foran DR. (2010) Bacterial profiling of soil using genus-specific markers and multidimensional scaling. *J Forensic Sci*. 55 (6):1437-42.

Macdonald CA, Ang R, Cordiner SJ, Horswell J. (2011) Discrimination of soils at regional and local levels using bacterial and fungal T-RFLP profiling. *J Forensic Sci*. 56(1): 61-9.

Marumo Y, Sugita R, Seta S. (1996) Soil as evidence in criminal investigation. 11th Interpol Forensic Science Symposium; Forensic Science Foundation Press.

Schabereiter-Gurtner C, Piñar G, Lubitz W, Rölleke S. (2001) Analysis of fungal communities on historical church window glass by denaturing gradient gel electrophoresis and phylogenetic 18S rDNA sequence analysis. *J Microbiol Methods*. 47 (3): 345-54.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 75 (23): 7537-41.