Metagenomic Characterization of The Entire Microbial Community in A Radiochemical Sample

Stefania Brandini, Alessio Valletti, Mariangela Santorsola, Antonio Riglietti, Anna Tolomeo, Onofrio Losito, Vincenzo Dimiccoli, Michele Diaferia, Sabino Liuni, Saverio Vicario, Elda Perlino

Abstract—**The detection of the mixed microbial communities in an environmental sample has represented a hurdle for long time due to the selection imposed by the culturing on specific media and by the minimum number of cells required. Here, we simulated a contamination of a sterile radiochemical by adding a known amount of *Staphylococcus epidermidis* and we developed a methodology, based on a metagenomic approach, to isolate DNA without a previous selection, from a low starting material and in limited time. The obtained sequences have been further used to characterize the microbial populations present in the sample using a bioinformatics tool.**

Index Terms—**16S rRNA gene, bioinformatics, metagenomics, NGS, single cell extraction**

I. Introduction

The capability to investigate environmentally and clinically significant microorganisms in a sample is a topic in constant development. Historically, sample culturing has represented the main method used to identify microorganisms, but it heavily depends on microbial specific cultures. Compared to classical laboratory procedures, culture-independent screening doesn't require bacterial growth on different media, thus avoiding an upstream selection. In detail, metagenomics usesfig a combination of DNA extraction, Next Generation Sequencing (NGS) and bioinformatics techniques to detect the whole microbiota in a sample, even starting from a very small number of cells. Considering that many microorganisms, particularly bacteria, can be pathogenic and are often the causative agent of diseases [1], their exact identification, even when present in small amount, can have a strong impact in clinical microbiology.

We developed a procedure that allows to detect and to identify bacteria starting from tiny sources. To obtain sufficient material for genetic analyses of DNA, we used the whole genome amplification (WGA) technique, in particular the isothermal method of multiple displacement amplification (MDA) [2, 3] that utilizes polymerases with high processivity and strand-displacement activity that extends from randomly primed sites [4].

Thanks to the availability of whole-genome sequences and taxonomic markers from microorganisms, it is possible to have a comprehensive overview of the whole microbiota in the sample [5] after the comparison of the obtained sequences with all available.

We applied this procedure to test the sterility of a new radiochemical (RC) produced at ITEL telcomunicazioni SRL corporation of Ruvo di Puglia, Italy, within the project 'Cluster in Bioimaging' (QZYCUM0). This RC is used for the diagnostics by positron emission tomography (PET) and it is characterized by a half-life of little more than one hour.

II. Results and discussion

In order to set up a useful protocol for the DNA extraction from our RC sample, we added a known amount of contaminants to the sterile matrix. A strain of *Staphylococcus epidermidis*, commonly present in the human skin flora, was inoculated in LB medium and the optical density at 600 nm (OD600) of the suspension was measured every 30 minutes to monitor the phases of cellular growth. This allowed us to predict the cell number of the contaminant by applying the formula 1 OD600 = 7 x 108 cell/ml, that is specific for *S. epidermidis* cells [6].

To verify the cell vitality, we prepared serial dilutions starting from an aliquot of the suspension taken during the linear phase of cellular growth (OD600 = 0.207, estimated cell/ml = 1.45 x 108). In this way, we also confirmed that the predicted number of cells from each dilution corresponded to the number of colonies obtained after overnight incubation on LB-agar plates (Table 1 and Fig 1).

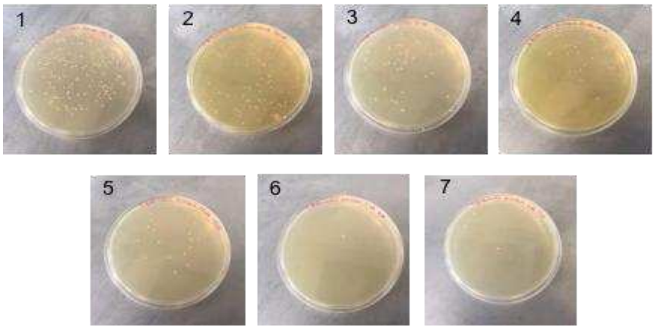

Figure 1: Validation of effective cell number and vitality. Number of colonies obtained after overnight (O/N) incubation of the LB-agar plates.

Stefania Brandini, Institute of Biomedical Technologies (ITB), National Research Council (CNR), Bari, Italy.
Alessio Valletti, ITB-CNR, Bari, Italy.
Mariangela Santorsola, ITB-CNR, Bari, Italy.
Antonio Riglietti, ITEL telcomunicazioni SRL, Ruvo di Puglia, Italy.
Anna Tolomeo, ITEL telcomunicazioni SRL, Ruvo di Puglia, Italy.
Onofrio Losito, ITEL telcomunicazioni SRL, Ruvo di Puglia, Italy.
Vincenzo Dimiccoli, ITEL telcomunicazioni SRL, Ruvo di Puglia, Italy.
Michele Diaferia, ITEL telcomunicazioni SRL, Ruvo di Puglia, Italy.
Sabino Liuni, ITB-CNR, Bari, Italy.
Saverio Vicario, Institute of Atmospheric Pollution Research (IIA), National Research Council (CNR), Bari, Italy.
Elda Perlino, ITB-CNR, Bari, Italy.

Table 1: Validation of effective cell number and vitality. Number of colonies obtained after overnight (O/N) incubation of the LB-agar plates compared with the number of colonies expected plating an equal volume of *S. epidermidis* suspension used for the DNA extraction.

Plate	*S. epidermidis* expected colonies	*S. epidermidis* colonies obtained after O/N incubation
1	150 expected cells	182 cells
2	100 expected cells	101 cells
3	50 expected cells	43 cells
4	25 expected cells	28 cells
5	15 expected cells	18 cells
6	5 expected cells	3 cells
7	1 expected cell	3 cells

After the validation of the cellular vitality, we tested the sensibility of the REPLI-g Single Cell Kit (Qiagen) extraction methods by isolating the genetic material starting from the theoretical number of 150, 100, 50, 25, 15, 5 and 1 cells of *S. epidermidis*. DNA quality was evaluated by gel electrophoresis (Fig 2).

In order to verify the bacterial origin of the extracted DNA, and not mere primer dimer, we carried out a PCR using primers specific for the prokaryotic target gene 16S rRNA, resulting in a 354 bp long product of the V5-V6 region.

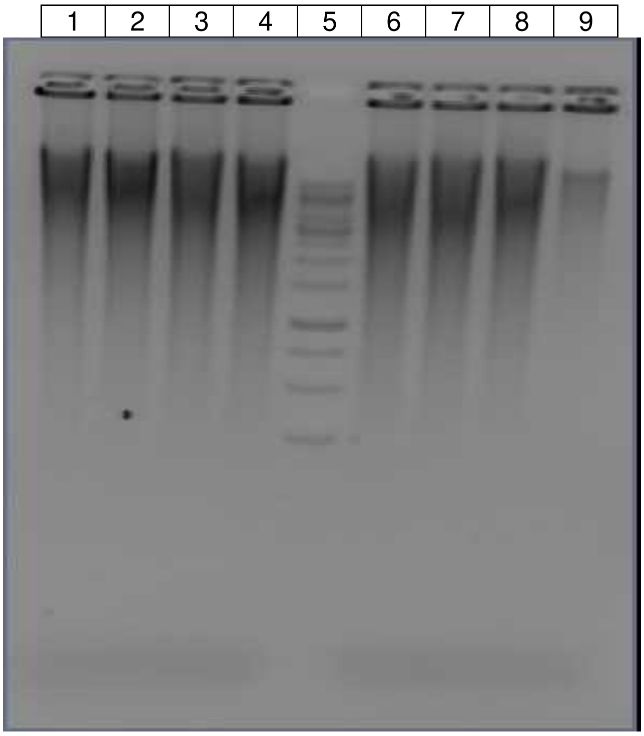

Fig 2: 1% agarose gel of DNA (5 µg) extracted starting from a theoretical number of:

- Lane 1 = 150 cells
- Lane 2 = 100 cells
- Lane 3 = 50 cells
- Lane 4 = 25 cells
- Lane 5 = 500 ng GeneRulerTM Ladder (Thermo Fisher 10000-250 bp)
- Lane 6 = 15 cells
- Lane 7 = 5 cells
- Lane 8 = 1 cell
- Lane 9 = 0 cells

The presence of products in the samples from 5 or more cells (Fig 3, Lane 2-7) confirms the extraction of DNA of bacterial origin, while their lack from the single cell (Fig 3, Lane 8) suggests a procedural difficulty in applying the protocol to a theoretical number of 1 cell.

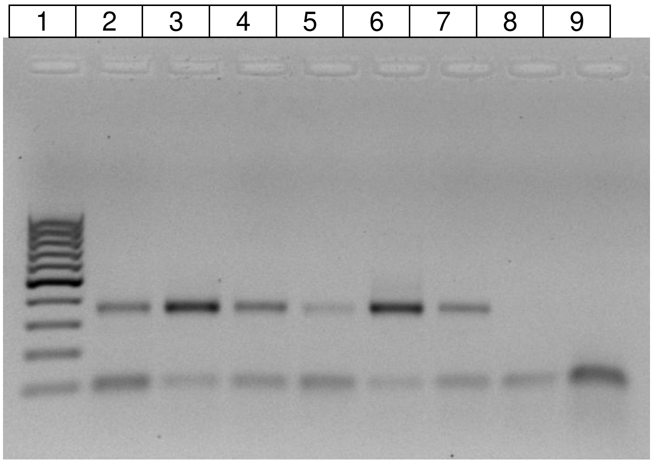

Fig 3: 1.5% agarose gel of amplified DNA from 1 ng of extracted DNA starting from:

- Lane 1 = 500 ng GeneRulerTM Ladder (Thermo Fisher 1000-100 bp)
- Lane 2 = 150 cells
- Lane 3 = 100 cells
- Lane 4 = 50 cells
- Lane 5 = 25 cells
- Lane 6 = 15 cells
- Lane 7 = 5 cells
- Lane 8 = 1 cells
- Lane 9 = 0 cells

All the further experiments have been leaded using 50 cells of contaminants, in order to avoid misrepresented results due to manual errors.

To identify the amplification condition in which the amount of extracted DNA linearly depended on the incubation times, we extracted DNA from 50 cells of *S. epidermidis* varying the time of the amplification step from eight to two hours at 30°C. The results demonstrated that the amount of DNA depends on the cell number when the amplification step of the extraction protocol proceeds for two hours (Fig 4, Lane 5). Finally, DNA was extracted in duplicate from both the RC and the PBS provided by the kit with and without the addition of 50 cells of *S. epidermidis* (Fig 5).

The presence of bacterial DNA was tested for all the RC samples by the amplification of the target V5-V6 region. As expected, we found bacterial DNA only in those samples where we added *S. epidermidis* (Fig 6, Lanes 4 and 5).

RC samples including bacterial DNAs underwent to the library preparation and sequencing using an Illumina MiSeq® platform as described in Methods.

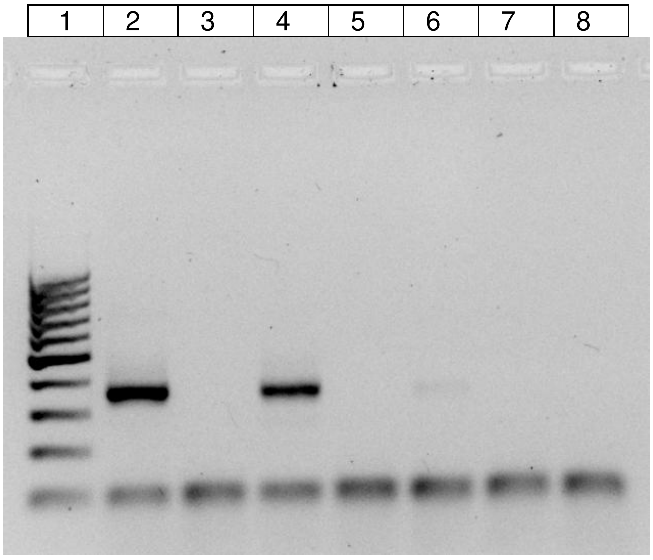

Figure 4: 1.5% agarose gel of amplified DNA from 1 ng of extracted DNA from:

- Lane 1 = 500 ng GeneRulerTM Ladder (Thermo Fisher 1000-100 bp)
- Lane 2 = sterilized water incubated for 8h (30° C)
- Lane 3 = RC incubated for 8h (30° C)
- Lane 4 = RC + 50 cells incubated for 8h (30° C)
- Lane 5 = RC incubated for 2h (30° C)
- Lane 6 = RC + 50 cells incubated for 2h (30° C)
- Lane 7 = sterilized water - polymerase incubated for 8h (30° C)
- Lane 8 = RC - polymerase incubated for 8h (30° C)

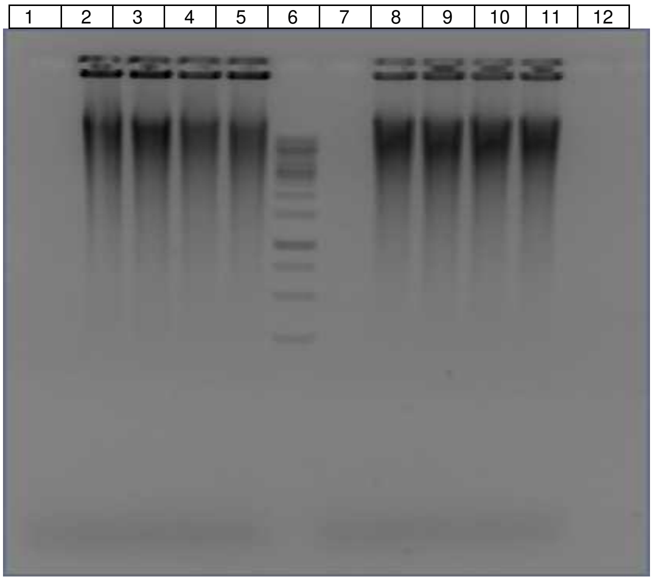

Figure 5: 1% agarose gel of DNA (5 µg) extracted from:

- Lane 1 = PBS - polymerase
- Lane 2 = PBS
- Lane 3 = PBS
- Lane 4 = RC
- Lane 5 = RC
- Lane 6 = 500 ng GeneRulerTM Ladder (Thermo Fisher 10000-250 bp)
- Lane 7 = PBS – polymerase + 50 cells
- Lane 8 = PBS + 50 cells
- Lane 9 = PBS + 50 cells
- Lane 10 = RC + 50 cells
- Lane 11 = RC + 50 cells
- Lane 12 = //

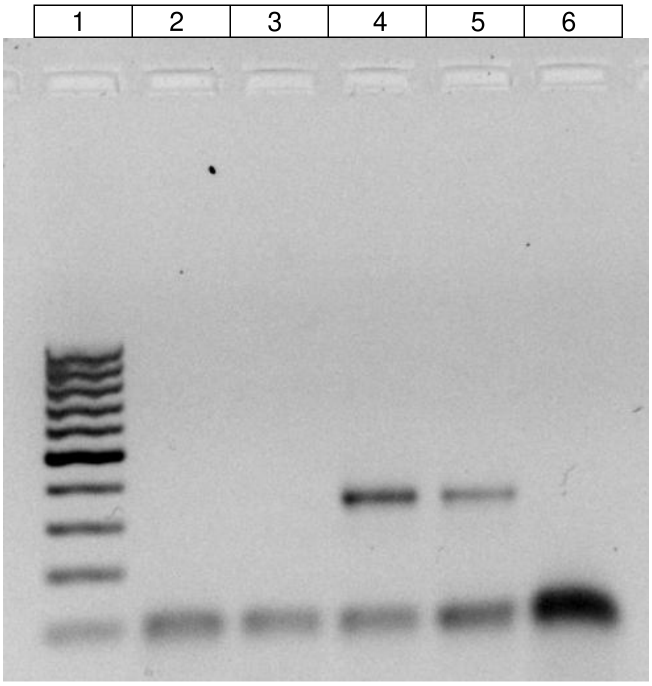

Figure 6: 1.5% agarose gel of amplified DNA from 1 ng of extracted DNA from:

- Lane 1 = 500 ng GeneRulerTM Ladder (Thermo Fisher 1000-100 bp)
- Lane 2 = RC
- Lane 3 = RC
- Lane 4 = RC + 50 cells
- Lane 5 = RC + 50 cells
- Lane 6 = negative control

The reads were filtered for noise (denoising) using cutadapt and vsearch. The tool cutadapt [7] removed primers and the ends of the reads with a quality lower than 20. Then, vsearch [8] allowed to merge the read pairs, to identify identical sequences (Unique) and to group them with the 98% of similarity (cluster) removing singleton reads (Table 2). Table 2 report the results of the denoising procedure on the sequences obtained following both the manufacturer's amplification condition at eight hours and reducing to two hours. Those results are comparable and suitable for NGS sequencing. The representativeness of the single sequences is reported in Table 3.

Table 2: Sequence parameters obtained during the denoising process.

	1=RC 8h	2=RC 8h	3=RC 8h	4=RC 2h	Overall
Total read pairs processed	219778	198189	213870	187123	
Merged	147937	92565	148987	124917	
Unique	9535	5649	9413	8045	
Clusters	1374	645	1510	1182	163
NoSingle	55	36	83	45	44

The different columns represents the 4 samples, while the rows indicate the different steps listed in material and methods. The overall column represent a clustering performed on the overall set of demultiplied sequences, in particular NoSingle in the overall columns represent sequence present in at least two sample with at least 4 reads.

Table 3: Count of corrected sequences.

#OTU ID	1=RC 8h	2=RC 8h	3=RC 8h	4=RC 2h
Read_1	147796	146671	116587	95989
Read_2	4594	3562	15896	578
Read_4	1841	2978	131	2261
Read_3	0	2862	837	291
Read_5	1891	621	561	616
Read_10	461	168	109	244
Read_8	434	8	143	232
Read_7	455	0	196	125
Read_13	180	124	76	224
Read_15	236	98	71	157
Read_6	500	0	0	29
Read_11	244	145	98	22
Read_9	0	229	0	228
Read_12	190	128	82	0
Read_14	14	99	65	169
Read_20	122	50	19	104
Read_16	0	97	41	154
Read_18	50	64	29	146
Read_25	95	33	16	128
Read_17	147	68	33	0

The 20 most numerous sequences among all the reads are reported. The total corrected sequences among the four experiments are 160.

The results of sequencing highlight the low variability among the sequences, in terms of small number of unique sequences (Table 2) and of a single leading taxonomic assignment (Table 3). This confirms the monoclonal nature of the bacterial sample of *S. epidermidis* used to contaminate the radiochemical.

Each query assignment, as the value of Likelihood Weight Ratio, representing the affinity between a sequence and a position in the reference tree, is reported in Table 4.

Table 4: Names and assignments of sequences by the pipeline Infernal, aEPA-NG and PAN.

Assignment[a]	Rank	N. of reads	Mean read assignment	Read %	Unique reads
Bacteria	superkingdom	508975	1.000008	99.89	16
Staphylococcus petrasii	species	507697	0.019370	99.65	13
Staphylococcus cohnii	species	507697	0.019370	99.65	13
Staphylococcus	genus	507697	0.932044	99.65	13
Staphylococcaceae	family	507697	0.999418	99.65	13
Staphylococcus sciuri	species	507697	0.038741	99.65	13
Salinicoccus	genus	507697	0.053971	99.65	13
Staphylococcus saprophyticus	species	507686	0.019370	99.65	12
Staphylococcus succinus	species	507686	0.019370	99.65	12
Pseudomonas fluorescens group	species group	1310	0.050450	0.25	5
Pseudomonas chlororaphis	species	1310	0.014256	0.25	5
Pseudomonas	genus	1310	0.825742	0.25	5
Pseudomonas syringae group	species group	1310	0.061327	0.25	5
Delftia	genus	121	0.166482	0.02	1
Diaphorobacter	genus	121	0.002510	0.02	1
Burkholderiales	order	121	1.000001	0.02	1
Hydrogenophaga	genus	121	0.121799	0.02	1
Comamonas	genus	121	0.041044	0.02	1
Acidovorax	genus	121	0.020262	0.02	1
Ottowia	genus	121	0.020119	0.02	1
Comamonadaceae	family	121	0.981764	0.02	1
Brachymonas	genus	121	0.040706	0.02	1
Simplicispira	genus	121	0.040768	0.02	1

Example obtained from an in vitro validation experiment using a monoclonal contamination by *S. epidermidis.*

From Table 4 is possible to grasp the distribution of the cluster sequences. Thirteen clusters over 44 represent more than 99% of the reads and all are assigned within Staphylococcaceae, with a support of 99.94%, and within Staphylococcus, with a support of 93.20%. The assignment to the species of Staphylococcus is very low (no more than 2%), consistent with the fact that *S. epidermidis* was present with a single sequence in LTPs128 (see Bioinformatics pipeline section) and no explicit assignment could have been done. Indeed, the homemade script PAN (Phylogenetic Assigner Namer) requires at least two sequences to assign a given taxonomic name. Two groups of reads tell possible alternative stories. Five clusters prefer Pseudomonas, while a single Cluster is assigned to the family Comamonadaceae. These two groups of assignment are probably caused by errors in the sequencing reaction and represent globally only 0.27% of the read pool.

This metagenomic approach confirmed the presence of *Staphylococcus* sequences in our samples (Table 4), but its application could be extended to the detection of all the contaminants present in a single sample, in order to verify its sterility and to classify all the potential communities within. In this context, the method here described would substitute the traditional microbiological tests crossing the selection step imposed by the sample culturing on a specific medium.

III. Materials and Methods

In order to be able to set up the best condition for DNA isolation we simulated a contamination of a sterile radiochemical (RC) by adding a known amount of *S. epidermidis*. For this purpose, the ITEL Telecomunicazioni SRL corporation of Ruvo di Puglia, Italy, supplied us with a radiochemical (RC) which overcame the quality tests of sterility provided for by the European normative, and with a strain of *S. epidermidis* captured in their hair filtering system. The sequences obtained after NGS sequencing have been successively used to set up the bioinformatics pipeline useful to characterize the whole microbial populations present in the starting sample.

A. Count of S. epidermidis cells

In order to contaminate the sterile radiochemical (RC) by a known amount of a bacterial species, we inoculated a strain of *S. epidermidis* in a LB medium and evaluated the number of cells in the suspension.

The number of cells has been estimated by measuring the optical density of the cell suspension at 600 nm (OD600) at the NanoDrop 2000UV-Vis and applying the formula 1 OD600 = 7 x 108 cell/ml [6]. In order to verify that the effective number of vital cells corresponded to the theoretical estimation, we counted the number of colonies on LB-agar plates after overnight incubation.

B. DNA extraction

DNA was extracted by REPLI-g Single Cell kit (Qiagen). The reactions were performed in 50 µl volume according to manufacturer's instructions. This kit allows the DNA extraction from a very low amount of starting material (even from a single cell) thanks to the Multiple Displacement Amplification (MDA) technology, which carries out an isothermal genome amplification utilizing the uniquely processive DNA polymerase Phi 29. The protocol includes a first step of cell lysis and DNA denaturation followed by an isothermal amplification reaction that proceeds for eight hours (h) at 30°C. These conditions were changed during the run-on experiment where the time of the amplification step was reduced from eight to four and two hours.

C. Sequencing

Bacterial DNA has been identified by the detection of the eubacterial 16S rRNA gene. Gene-specific sequences targeting the V5-V6 hypervariable regions of the 16S gene were amplified as described in [9], using the primer pairs V5-V6 NextFor/Rev. 16S rRNA gene amplicon sequencing was carried out to identify the bacterial profiles present in the samples. For the preparation of paired-end metagenomic DNA libraries, the PCR products have been purified by AMPure XP Beads and undergone to indexing following the Nextera DNA sample preparation guide (Illumina Incorporate, California USA). Purified amplicons were pooled and sequenced on an Illumina MiSeq®.

D. Bioinformatics pipeline

First, primer sequences and read end bases with a quality lower than 20 were removed with cutadapt. Next, sequence pairs were merged and then demultipled, namely clustered with 100% identity. Following, the read abundance of demultiplication step was used as guide by vsearch to cluster with 98% similarity threshold. The longest and most abundant read of each cluster was used as representative.

A reference sequence alignment was built from the bacterial section of the Living Tree version 128 of Silva database (LTPs128) [10]. Specifically, these sequences were realigned using Infernal [11] with Rfam seed alignment for SSU rRNA (Small Subunit Ribosomal) profile. The alignment was modified only marginally just to accommodate the SSU rRNA sequences of Rfam not included in SILVA alignment. These allowed to add the secondary structure annotation of Rfam and to build the Hidden Markov Model of the alignment that allowed fast addition of the query sequences. Then, cluster representatives (queries), were aligned to the new reference alignment by Infernal. The queries were placed on the phylogenetic tree of LTPs128 using the EPA-NG implementation (https://github.com/Pbdas/epa-ng) of the phylogenetic placement algorithm [12]. This new implementation, differently from the version implemented in RaxML, registers more suboptimal placement for each query. This meets the need to be conservative, given the necessity to build a monitoring tool to identify any contamination.

The phylogenetic placement information was translated in a taxonomic report using as reference NCBI taxonomy and the new script PAN (Phylogenetic Assigner Namer). The script works in three steps. In the first one it works only on the output of the phylogenetic assigner. For each query, it identifies the direct ancestor node of each query insertion point and assigns its relative likelihood score. Next, the entire subtree defined by the most recent common ancestor of all query insertion points is traversed in post order adding up to each encountered node the score of all its descendants. In the second step labels from taxonomy are added to each internal node and relevant rank are identified. The NCBI lineage string is added to each terminal node of the LTPs128 phylogenetic tree. Always traversing in post order the tree, the most lower common taxonomic rank of all descendant leaves of a given nodes is used as taxonomic label. If a direct ancestor node has the same label than one of its direct descendant the label is erased in the descendant. This ensures that labeled nodes are the deepest in the tree and they should collect as much relative likelihood as possible from descendant. In the third step a summary taxonomic report is built. The mean of the likelihood scores and the total number of reads supporting a given label is returned across assigned query (Fig 7).

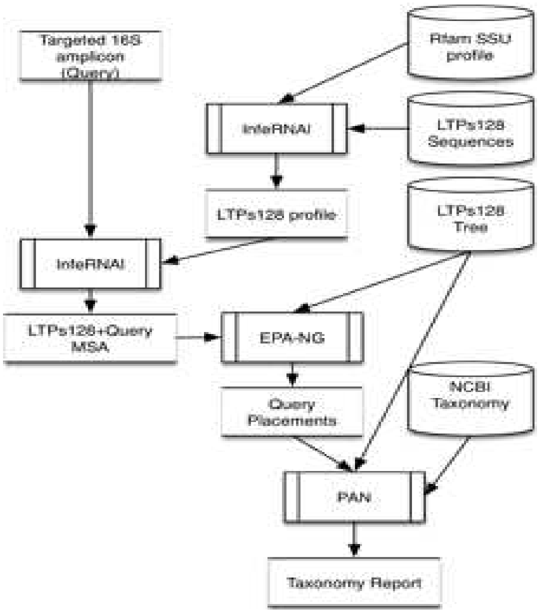

Figure 7: Workflow of the pipeline PAN (Phylogenetic Assignment Namer) to obtain the taxonomic assignment of the sequences deriving from 16S RNA amplicons.

ACKNOWLEDGMENT

This work was supported by the project 'Cluster in Bioimaging' (QZYCUM0) by the Apulian Region. The authors wish to thank Giuseppe Cananzi for technical assistance and Dr Caterina Manzari for the valuable suggestions.

REFERENCES

[1] Wright MH, Adelskov J and Greene AC. "Bacterial DNA extraction using individual enzymes and phenol/chloroform separation." J. Microbiol. Biol. Educ. 2017. 2.

[2] Blainey PC. "The future is now: single-cell genomics of bacteria and archaea." FEMS Microbiol. Rev. 2013. 3: 407-427.

[3] Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M and Lasken RS. "Comprehensive human genome amplification using multiple displacement amplification." Proc. Natl. Acad. Sci. U S A. 2002. 8: 5261-5266.

[4] de Bourcy CF, de Vlaminck I, Kanbar JN, Wang J, Gawad C and Quake SR. "A quantitative comparison of single-cell whole genome amplification methods." PLoSOne. 2014. 8: e105585.

[5] Leo S, Gaïa N, Ruppé E, Emonet S, Girard M, Lazarevic V and Schrenzel J. "Detection of bacterial pathogens from broncho-alveolar lavage by Next-Generation Sequencing." Int. J. Mol. Sci. 2017. 9: 189.

[6] Unciti-Broceta JD, Cano-Cortés V, Altea-Manzano P, Pernagallo S, Díaz-Mochón JJ and Sánchez-Martín RM. "Number of nanoparticles per cell through a spectrophotometric method - a key parameter to assess nanoparticle-based cellular assays." Sci. Rep. 2015. 5: 10091.

[7] Martin M. "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet.journal. 2011. 1: 10-12.

[8] Rognes T, Flouri T, Nichols B, Quince C and Mahé F. "VSEARCH: a versatile open source tool for metagenomics." PeerJ. 2016. 4: e2584.

[9] Manzari C, Fosso B, Marzano M, Annese A, Caprioli R, D'Erchia AM, Gissi C, Intranuovo M, Picardi E, Santamaria M, Scorrano S, Sgaramella G, Stabili L, Piraino S. and Pesole G. "The influence of invasive jellyfish blooms in a coastal lagoon (Varano, SE Italy) detected by an Illumina-based deep sequencing strategy." Biol. Invasions. 2015. 17: 923–940.

[10] Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO and Rosselló-Móra R. "Release LTPs104 of the All-Species Living Tree." Syst. Appl. Microbiol. 2011. 34: 169-170.

[11] Nawrocki EP and Eddy SR. "Infernal 1.1: 100-fold faster RNA homology searches." Bioinformatics. 2013. 22: 2933-2935.

[12] Berger SA, Krompass D and Stamatakis A. "Performance, accuracy, and Web server forevolutionary placement of short sequence reads under maximum likelihood." Syst. Biol. 2011. 3: 291-302.