Bacterial identification through accurate library preparation and high-throughput sequencing

Ana Paula Christoff[1]; Aline Fernanda Rodrigues Sereia[1]; Dellyana Rodrigues Boberg[1]; Rômulo Lucio Vale de Moraes[1] and Luiz Felipe Valter de Oliveira[1*]

[1]**Neoprospecta Microbiome Technologies, SA**
Address correspondence to: Luiz Felipe Valter de Oliveira, Neoprospecta Microbiome Technologies,
Av. Luiz Boiteux Piazza, 1302, Sapiens Parque, Florianópolis, SC.
Email: felipe@neoprospecta.com

Highlights

Bacterial high-throughput screening for diverse, complex and low input samples;

High-sensitive detection of bacterial DNA in hospital, food or natural environments;

Accurate 16S rRNA gene (V3-V4) library preparation for Illumina sequencing;

Keywords
Bacteria; next-generation sequencing; high-throughput sequencing; 16S rRNA, NGS..

Summary

Large scale DNA sequencing can provide unprecedented information about bacterial composition in a sample. A high-sensitive screening of 16S rRNA gene was optimized for diverse, complex and low input samples. Using a two-step PCR we evaluated the most suitable library preparation method for high-throughput Illumina sequencing, allowing bacterial screening in hospital environments, food industries and products, pharmaceuticals or natural environments. To achieve great accuracy in the results, 16S rRNA gene V3-V4 hypervariable region was amplified with standard hot start DNA polymerase. Such library preparation together with a stringent bioinformatics analysis, to remove chimeric and erroneous reads, allow high sensibility in bacterial identification from complex and low input samples. Additionally, other gene markers could be used in the same workflow to gain resolution about specific microorganisms. These results can be very helpful in understanding the whole microbiota present in a sample and their associations with ecological balance for human and environmental health.

New sequencing methodologies are providing fast and high-resolution results, revolutionizing the microbiology field. Countless microorganisms that could not be cultivated so far, can now be deeply characterized using molecular techniques based on DNA sequences. The clinical microbilogy field, food industries, pharmaceuticals or anyone focused on rapidly characterize a pathogenic microorganism can benefit from these acurate methodology of DNA sequencing. Broadly used to identify bacteria, the 16S rRNA gene is actually the most comprehensive marker gene available, since it has good phylogenetic resolution among bacteria, and one of the largest DNA sequence databases (Hugenholtz et al., 2016; Yang et al., 2016). A great advantage of 16S rRNA gene sequencing is the possibility of extensive bacterial screening directly from environmental samples, without the need for microorganism isolation. In fact, this is the main goal of our library preparation and sequencing method: to perform sensitive bacterial large scale screening and identify these microorganisms in highly diversified set of samples.

16S rRNA gene contains nine hypervariable regions (V1-V9) that can be used for taxonomy studies (Yang et al., 2016). The accuracy in 16S rRNA gene identification is directly dependent on several factors concerning library preparation, DNA sequencing and data analysis (Tremblay et al., 2015; Kebschull et al., 2015; Gohl et al., 2016). Several protocols have been described, including one widely used from Earth Microbiome Project (http://www.earthmicrobiome.org/protocols-and-standards/16s/), or the 16S metagenomic sequencing library preparation from Illumina (Illumina Technical Note 15044223 Rev. B). Nevertheless, most of these protocols require high amounts of DNA (e.g. 12.5ng) to amplification with high-fidelity enzymes. In many environmental samples, as we described here, this amount of DNA could not be obtained. So, we developed a highly sensitive library preparation method, comparing the standard amplification of V4 16S rRNA region with our optimized V3-V4 region. In addition, we compared our method with the Illumina's standard protocol, which uses high-fidelity polymerase.

Following the best practices of sample collection, with appropriate storage and transport conditions to preserve the bacterial DNA, we can utilize DNA from basically any extraction method if it results in good quality DNA, even in low amounts (less than 0.05ng). Our sequencing library preparation was carried out in a two-step PCR protocol. This has been shown as advantageous (Gohl et al., 2016), giving the better primer amplification efficiencies and optimization for multiplexing. In the first PCR reaction, we used the V3-V4 primers 314F-806R (Wang et al., 2009; Caporaso et al., 2011), since this pair has great taxonomy coverage in bacteria and archaea (Takahashi et al., 2014). The commonly used pair 515F-806R (Caporaso et al., 2011) from Earth Microbiome Project was additionally tested, but it showed less sensitivity to discriminate more close related species (Fig 1A).

Then, in our protocol we choose the V3-V4 primers with the following conditions: the first PCR primers contain the Illumina sequences based on TruSeq structure adapter (Illumina, San Diego, CA), allowing the second PCR with indexing sequences. The PCR reactions were always carried out in triplicates using Platinum Taq (Invitrogen, USA) with the conditions: 95°C for 5 min, 25 cycles of 95°C for 45s, 55°C for 30s and 72°C for 45s and a final extension of 72°C for 2 min for PCR 1. In PCR 2 the conditions were 95°C for 5 min, 10 cycles of 95°C for 45s, 66°C for 30s and 72°C for 45s and a final extension of 72°C for 2 min. For comparison, Illumina 16S protocol was used as described (Illumina Technical Note 15044223 Rev. B). The final PCR reaction was cleaned up using AMPureXP beads (Beckman Coulter, Brea, CA) and samples were pooled in the sequencing libraries for quantification. Amplicon estimations were performed with Picogreen dsDNA assays (Invitrogen, USA), and then the libraries were diluted for accurate qPCR quantification using KAPA Library Quantification Kit for Illumina platforms (KAPA Biosystems, Woburn, MA).

The libraries were sequenced in a MiSeq system, using the standard Illumina primers provided in the kit. Usually a single-end 300nt run was performed.

After sequencing, our bioinformatics pipeline performs sequence demultiplexing, adaptor and primer trimming. The reads were size normalized to 283pb.

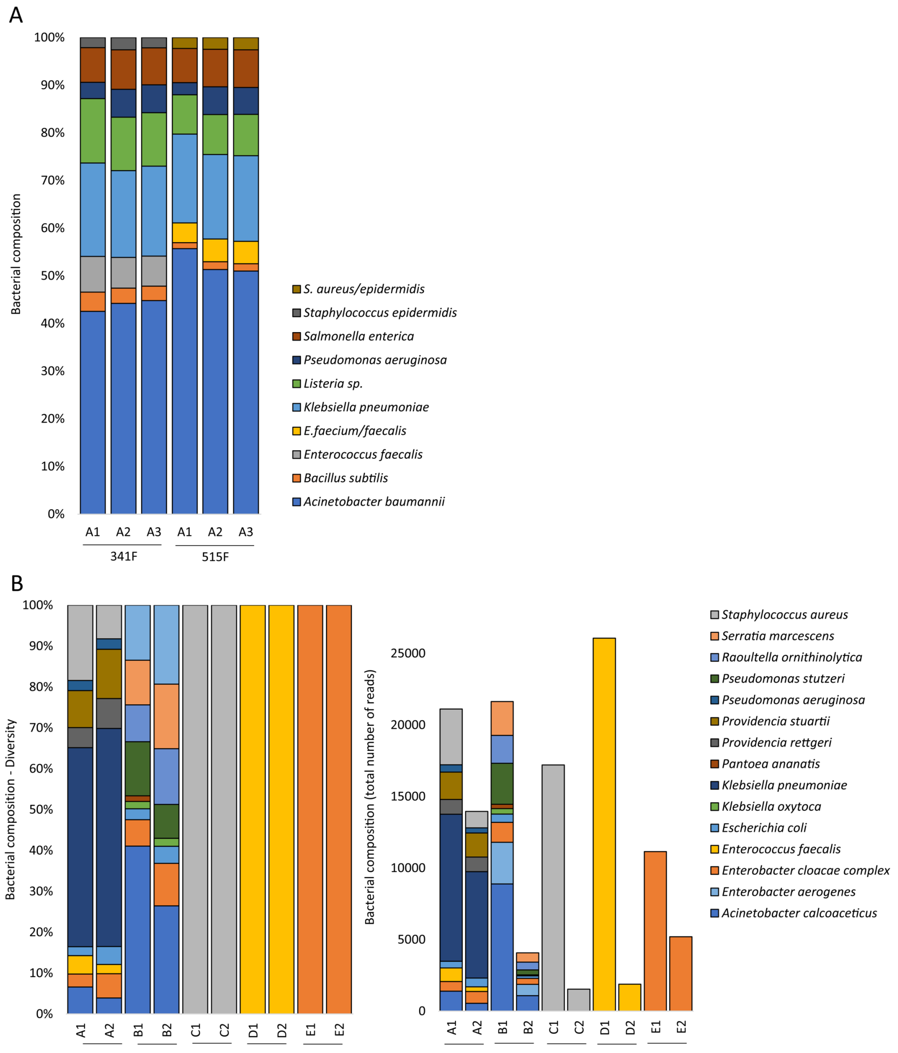

Figure 1. Bacterial profiles for 16S rRNA gene amplification with different primers and enzymes in mock samples. (A) Hypervariable regions V3-V4 and V4 were amplified with the primers 341F-806R and 515F-806R in triplicates of the same sample. The region V4 alone shows less resolution to classify some species such as *Staphylococcus aureus* and *Enterococcus faecalis*. **(B)** Amplification with Taq or Hifi enzyme recovers the same diversity profiles (left), however high fidelity enzymes have higher requirements of DNA input to result in more sequenced reads (quantitative view in right). All reactions with Platinum Taq (A1, B1, C1, D1 and E1) and KAPA Hifi (A2, B2, C2, D2 and E2) were performed simultaneously with the same DNA input of pools with known bacterial species.

Read quality filter (E) was performed converting each nucleotide Q score in error probability (e_i), that was summed and divided by read length (L):

$$e_i = 10^{\frac{-Q_i}{10}} \qquad\qquad E = \frac{\sum_{i=1}^{n} e_i}{L}$$

If E was minor or equal to 0,01 (1%) the read was considered in downstream analysis.

To increase the reliability of the read, excluding possible diversity generated by chimeric amplicons or erroneous nucleotide incorporated in PCR, we cluster 100% identical reads. If any cluster is represented by fewer than 5 reads, it is not considered in further analysis.

In our pipeline, each cluster gets a unique identifier, allowing traceability among results. This allow us to compare OTUs (Operational Taxonomic Units) from different experiments, comparing taxonomies with different pipelines and reference database analysis. Using this approach of read clustering we stablished that our cluster is the OTU.

Clustered sequences (OTUs) were then subjected to taxonomic classification comparing them with our 16S rRNA database (NeoRefdb, Neoprospecta Microbiome Technologies, Brazil). Sequences with at least 99% of identity in the reference database will be taxonomically assigned.

Once we defined that V3-V4 primer is better suitable for 16S rRNA amplification, given its level of species resolution, we further achieved the best results using standard hot start polymerase in the library preparation, compared to Illumina's protocol using high-fidelity (Figure 1B). With high-fidelity, we observed the same bacterial profile and read sequences as in Platinum reactions (Fig 1B right). However, with high-fidelity we were unable to recover bacterial sequences from samples with fewer microorganisms, often resulting in false negatives. Additionally, high fidelity reactions often have lower read counts (Fig 1B, left), sometimes loosing rare species, as *Pantoea ananatis* in sample B. Our protocol showed higher sensibility for bacterial identification in a complexity of samples. Then, we established that Platinum Taq, along with stringent bioinformatics pipeline analysis, can be used to generate reliable amplicon libraries.

Figure 2 shows the high reproducibility of our library preparation and sequencing method for several natures of samples, including **microbiomes** (Fig 2A), communities of isolated bacteria **(mocks)** (Fig 2B), **food** products (Fig 2C) and **hospital** environments (Fig 2D). In the great majority, these samples have specific DNA amounts lower than 5ng.

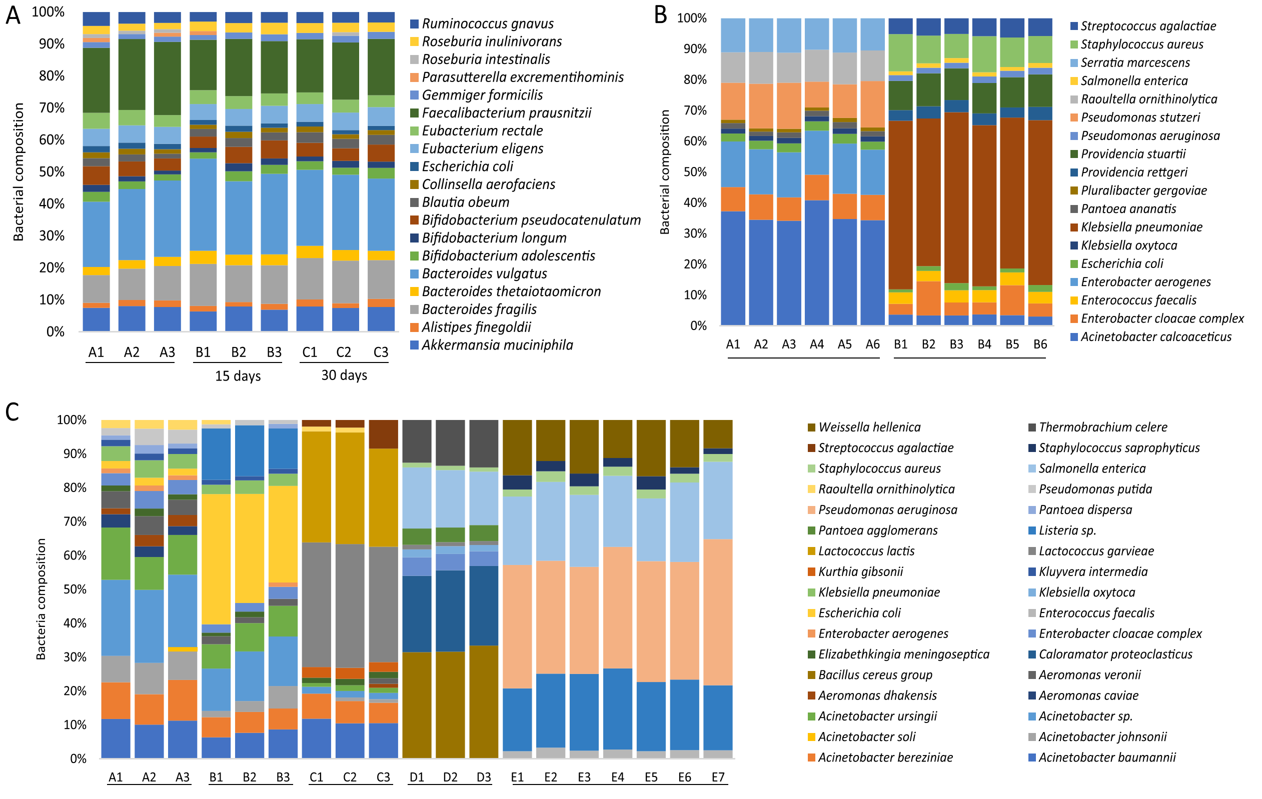

Figure 2. **Reproducible library and sequencing profiles in different types of samples. (A) Human fecal microbiome** sample analyzed after 15 and 30 days of sample storage. 1, 2 and 3 represent technical replicates of DNA extraction and library preparation for the same sample. **(B) Two bacterial communities** of culture-isolated DNA (A and B) went six replications of library preparation and sequencing. **(C) Food samples** were more diverse in bacterial composition, but 16S rRNA gene screening through our library preparation still shows great bacterial recovery. Samples A, B, C and D were analyzed in triplicates, whereas sample E was analyzed in seven replicates of DNA extraction, library preparation and sequencing. Continue.

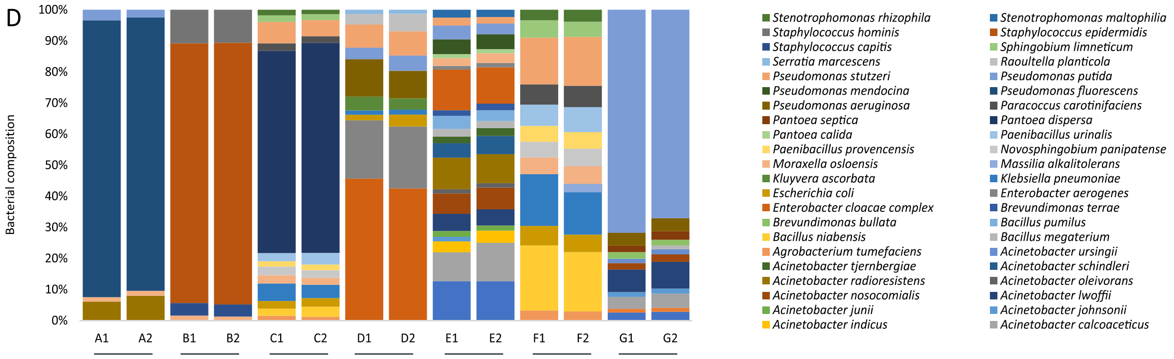

Figure 2 continuation. **(D) Hospital environments** are among the most difficult ones, since they have reduced amounts of bacteria. However, our library sequencing method shows high reproducibility among the bacterial composition recovery in different samples (A, B, C, D, E, F and G), prepared and sequenced in duplicates.

Conclusion

Our accurate method and analysis pipeline allows confidence in sequenced reads, not considering sequencing errors, chimeras or PCR artifacts in the bacterial taxonomic evaluation. Additionally, we can apply this protocol to a broad range of gene markers such as rpoB, gyrB, species specific virulence/pathogenicity genes or even antibiotic resistance genes (Lan et al., 2016; Law et al., 2015) to gain a better resolution about the taxonomies and characteristics of species found in 16S rDNA screened samples.

Acknowledgements

We would like to thank to Neoprospecta team members who helped in laboratory experiments Caroline Salvati, Laís Yamanaka, Pâmela Abdala and Patrícia Cunha, as well as in bioinformatics analysis Guilherme Cesar.

References

Caporaso, J., Lauber, C., Walters, W., Berg-Lyons, D., Lozupone, C., Turnbaugh, P., Fierer, N., Knight, R., n.d. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences 108 Suppl, 4516.

Gohl, D., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T., Clayton, J., Johnson, T., Hunter, R., Knights, D., Beckman, K., 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 34, 942–949.

Hugenholtz, P., Skarshewski, A., Parks, D., 2016. Genome-Based Microbial Taxonomy Coming of Age. Csh Perspect Biol 8, a018085.

Illumina 16S metagenomic sequencing library preparation (Illumina Technical Note 15044223 Rev. B). Illumina https://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf (2017).

Kebschull, J., Zador, A., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res 43, e143.

Lan, Y., Rosen, G., Hershberg, R., 2016. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. Microbiome 4, 18.

Law, J. W.-F., Mutalib N.-S. A., Chan K.-G., Lee, L.-H., 2015. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front Microbiol 5.

Takahashi, S., Tomita, J., Nishioka, K., Hisada, T., Nishijima, M., 2014. Development of a Prokaryotic Universal Primer for Simultaneous Analysis of Bacteria and Archaea Using Next-Generation Sequencing. Plos One 9, e105592.

Tremblay, J., Singh, K., Fern, A., Kirton, E., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J., Tringe, S., 2015. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in Microbiology 6.

Yang, B., Wang, Y., Qian, P.-Y., 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. Bmc Bioinformatics 17, 135.

Wang, Y., Quian, P-Y., 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PloS one 4, e740].