Improved metagenomic analysis using low error amplicon sequencing libraries

Ovidiu Rücker*, Alexandra Dangel*, Stefan Kotschote

IMGM Laboratories GmbH, Martinsried, Germany

Metagenomic analyses using 16S rRNA or other amplicons are widespread and performed on several next generation sequencing (NGS) platforms. During amplification or library preparation, several errors are inserted by polymerases and can occur subsequently during sequencing. Unlike resequencing approaches, metagenomic analyses lack a correcting reference genome, thus such errors cannot be detected or corrected.

Amplicon Libraries accumulate errors through multiple amplification steps:

- Amplicon generation during 30 or more cycles of PCR amplification (Error rate varies with polymerase type)
- Library amplification prior to sequencing:
  - Up to 50 cycles of emPCR amplification (Roche and Ion Torrent)
  - 10 cycles of bridge amplification (Illumina)
- Sequencing errors (Error rate depends on the sequencing technology)

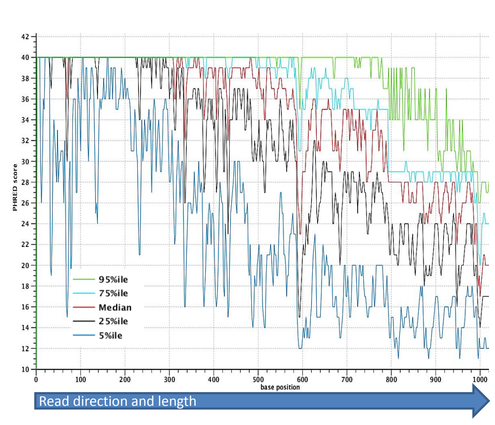
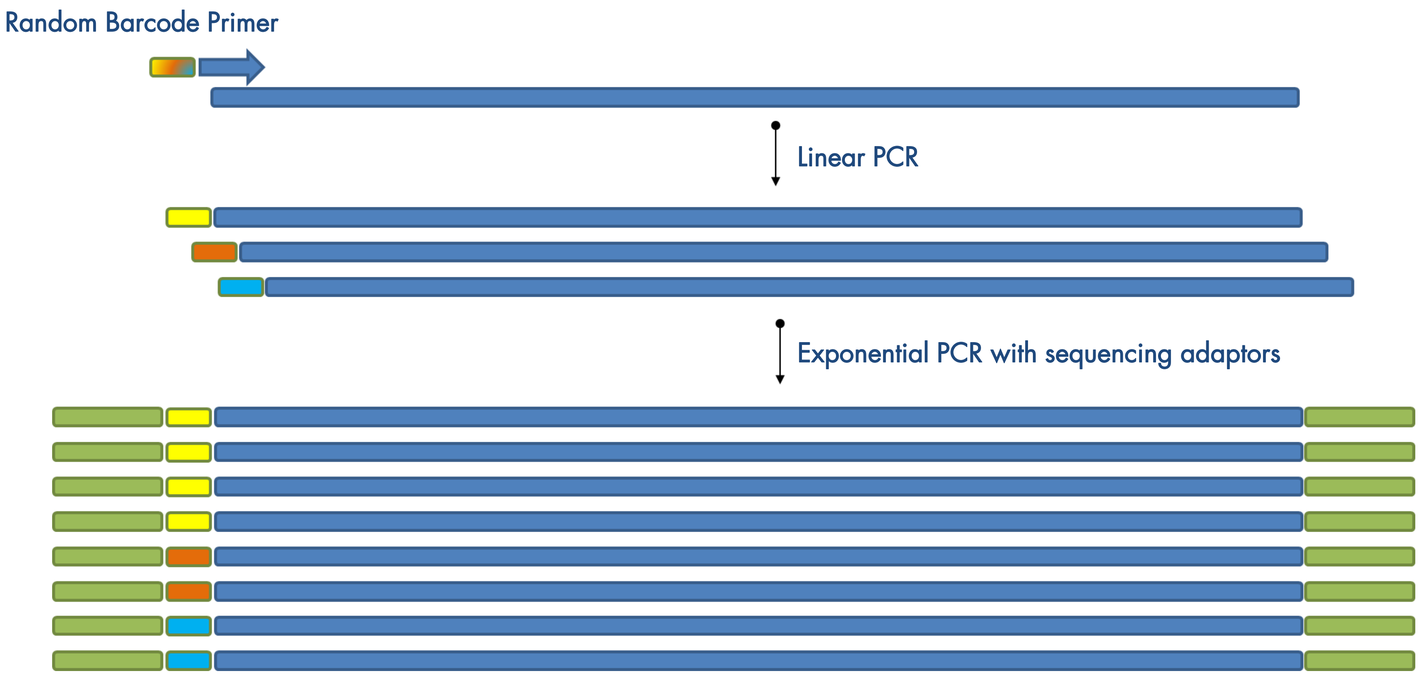


Fig. 1: Quality distribution along the read length (GS-FLX+, Roche). Phred score qualities of the median are indicated in red. The 95th, 75th, 25th and 5th percentile are also depicted using different colors indicated in the legend

Library Preparation for LEA-Seq

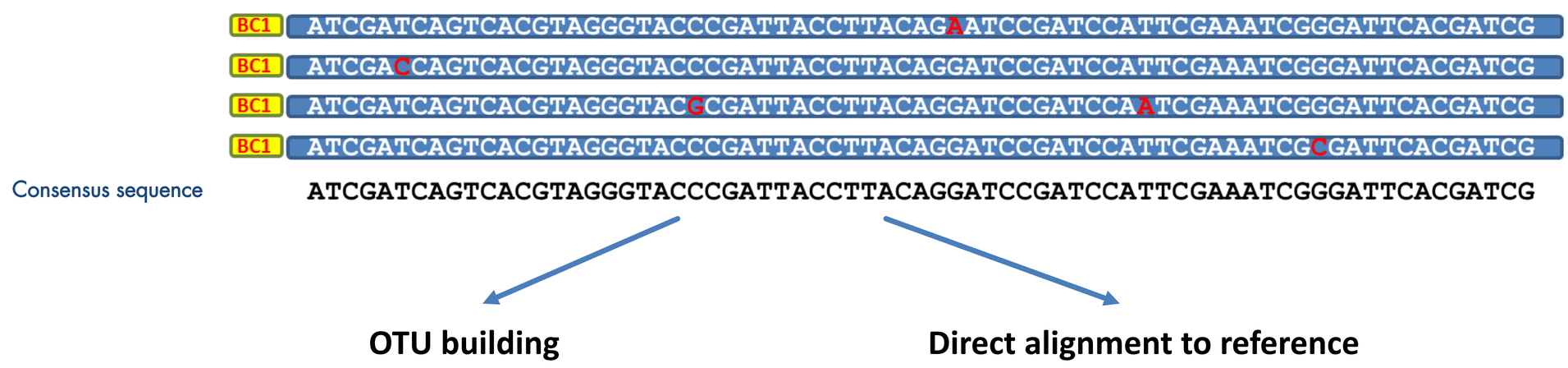


- Libraries comprising amplicons derived from a mock community of species with different habitats (e.g. soil, marine, human associated) were used for generating amplicon libraries
- We have inserted unique barcodes using random sequences. A linear PCR step yielded a low amount of initial amplification products. (Fig.2)
- These templates were then exponentially amplified and sequencing adaptors were inserted for several NGS platforms.

Fig. 2: Graphic overview of the amplicon library generation for the LEA-Seq method. During a linear PCR step random barcodes are introduced to the template. The following exponential PCR introduces the sequencing adaptors enabling sequencing with various NGS platforms.

Error correction during bioinformatic analysis

Using random barcodes it is feasible to pool all sequences derived from one initial DNA template (which now have the same unique barcode) together and build one consensus sequence. In our experiments we have observed an even distribution of cluster sizes, but also single clusters with more than 50 reads. In this way most errors which occur after the linear elongation can be corrected using bioinformatics tools. Duplicate sequences within clusters are also eliminated in this step, thus diminishing PCR amplification bias.



Conclusions

- This combination of library preparation and bioinformatic analysis (according to (1)) improves the quality of NGS data, by enabling the correction of random errors which occur before or during sequencing
- The elimination of such amplification errors directly leads to better phylogenetic assignment in metagenomic projects
- The generation of such low error amplicon reads enables a more precise insight into the analyzed community

*) both authors contributed equally

IMGM Laboratories GmbH

Bunsenstr. 7a | 82152 Martinsried, Germany | ph.: +49.89.4524667.0 | fax: +49.89.4524667.410 | info@imgm.com

1. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, u. a. The Long-Term Stability of the Human Gut Microbiota. Science. 7. Mai 2013;341(6141):1237439.