Tools for Quality Control and Preprocessing of Metagenomic Datasets

Robert Schmieder\textsuperscript{1,2}, Yan Wei Lim\textsuperscript{3} and Robert Edwards\textsuperscript{2,4}

\textsuperscript{1} Computational Science Research Center, SDSU; \textsuperscript{2} Department of Computer Science, SDSU; \textsuperscript{3} Department of Biology, SDSU; \textsuperscript{4} Mathematics and Computer Science Division, ANL

Summary of the computational processes that were integrated into a pipeline for automatic analysis. The web-based version allows the data processing through a user-friendly interface.

In 145 (72%) metagenomes, at least one possible contamination sequence was found using a threshold of 95% query coverage and 94% alignment identity.

• Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample
• Sequence contaminations are a serious concern to the quality of the data used for downstream analysis

Length distribution: Reads should be approximately the same length (same number of cycles). Short reads are likely lower quality reads.

Quality scores: Linearly degrading quality across the reads may require quality trimming as most assemblers or aligners do not take into account quality scores. Errors in low quality reads complicate assembly, might cause misassembly, or make assembly impossible.

Sequence duplications: Artificial duplicates can result in false variant (SNP) calling, wrong abundance or expression measures, and require more computing resources.

Additional features include GC content distribution; occurrence of ambiguous bases, poly-A/T tails and tag sequences; sequence complexity; assembly quality measures; and environmental classification using PCA on dinucleotide odds ratios.

Features for quality control and preprocessing

- Downstream sequence analysis is often compromised by low-quality sequences, sequence artifacts and sequence contamination, eventually leading to misassembly and erroneous conclusion

Features for quality control and preprocessing

- Metagenomes that were pre-amplified with primer-based methods require the removal of additional tag sequences
- Sequenced reads can contain deletions or insertions due to sequencing limitations, and primer sequences may contain ambiguous bases

Example data showing nucleotide frequencies (top), predicted tag sequence (middle), and frequency range and median (bottom) for the first 50 positions with a clear separation between the non-random nucleotide positions of the tag (A), the quasi-random nucleotides of the tag (B) and the metagenomic sequence (C).

Simplified model showing how fragment-to-fragment concatenations can generate artificial concatenated sequences (left). Example of imperfect primer annealing that causes mismatch-induced mutations in the sequence reads (top).

Nucleotide frequency logos provided on the web interface that show the raw frequencies (top) and the filtered and corrected frequencies (bottom). The tag sequence can be identified more easily using filtered frequencies and allows more accurate predictions.

Features for quality control and preprocessing

- Sequence duplications: Artificial duplicates can result in false variant (SNP) calling, wrong abundance or expression measures, and require more computing resources.

Features for quality control and preprocessing

- Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample
- Sequence contaminations are a serious concern to the quality of the data used for downstream analysis

Features for quality control and preprocessing

- Metagenomes that were pre-amplified with primer-based methods require the removal of additional tag sequences
- Sequenced reads can contain deletions or insertions due to sequencing limitations, and primer sequences may contain ambiguous bases

Example data showing nucleotide frequencies (top), predicted tag sequence (middle), and frequency range and median (bottom) for the first 50 positions with a clear separation between the non-random nucleotide positions of the tag (A), the quasi-random nucleotides of the tag (B) and the metagenomic sequence (C).

Simplified model showing how fragment-to-fragment concatenations can generate artificial concatenated sequences (left). Example of imperfect primer annealing that causes mismatch-induced mutations in the sequence reads (top).

Nucleotide frequency logos provided on the web interface that show the raw frequencies (top) and the filtered and corrected frequencies (bottom). The tag sequence can be identified more easily using filtered frequencies and allows more accurate predictions.

Features for quality control and preprocessing

- Additional features include GC content distribution; occurrence of ambiguous bases, poly-A/T tails and tag sequences; sequence complexity; assembly quality measures; and environmental classification using PCA on dinucleotide odds ratios.

Features for quality control and preprocessing

- Sequence duplications: Artificial duplicates can result in false variant (SNP) calling, wrong abundance or expression measures, and require more computing resources.

Features for quality control and preprocessing

- Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample
- Sequence contaminations are a serious concern to the quality of the data used for downstream analysis

Features for quality control and preprocessing

- Metagenomes that were pre-amplified with primer-based methods require the removal of additional tag sequences
- Sequenced reads can contain deletions or insertions due to sequencing limitations, and primer sequences may contain ambiguous bases

Example data showing nucleotide frequencies (top), predicted tag sequence (middle), and frequency range and median (bottom) for the first 50 positions with a clear separation between the non-random nucleotide positions of the tag (A), the quasi-random nucleotides of the tag (B) and the metagenomic sequence (C).

Simplified model showing how fragment-to-fragment concatenations can generate artificial concatenated sequences (left). Example of imperfect primer annealing that causes mismatch-induced mutations in the sequence reads (top).

Nucleotide frequency logos provided on the web interface that show the raw frequencies (top) and the filtered and corrected frequencies (bottom). The tag sequence can be identified more easily using filtered frequencies and allows more accurate predictions.

Features for quality control and preprocessing

- Sequence duplications: Artificial duplicates can result in false variant (SNP) calling, wrong abundance or expression measures, and require more computing resources.

Features for quality control and preprocessing

- Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample
- Sequence contaminations are a serious concern to the quality of the data used for downstream analysis

Features for quality control and preprocessing

- Metagenomes that were pre-amplified with primer-based methods require the removal of additional tag sequences
- Sequenced reads can contain deletions or insertions due to sequencing limitations, and primer sequences may contain ambiguous bases

Example data showing nucleotide frequencies (top), predicted tag sequence (middle), and frequency range and median (bottom) for the first 50 positions with a clear separation between the non-random nucleotide positions of the tag (A), the quasi-random nucleotides of the tag (B) and the metagenomic sequence (C).

Simplified model showing how fragment-to-fragment concatenations can generate artificial concatenated sequences (left). Example of imperfect primer annealing that causes mismatch-induced mutations in the sequence reads (top).

Nucleotide frequency logos provided on the web interface that show the raw frequencies (top) and the filtered and corrected frequencies (bottom). The tag sequence can be identified more easily using filtered frequencies and allows more accurate predictions.

Features for quality control and preprocessing

- Sequence duplications: Artificial duplicates can result in false variant (SNP) calling, wrong abundance or expression measures, and require more computing resources.

Features for quality control and preprocessing

- Sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample
- Sequence contaminations are a serious concern to the quality of the data used for downstream analysis

Features for quality control and preprocessing

- Metagenomes that were pre-amplified with primer-based methods require the removal of additional tag sequences
- Sequenced reads can contain deletions or insertions due to sequencing limitations, and primer sequences may contain ambiguous bases

Example data showing nucleotide frequencies (top), predicted tag sequence (middle), and frequency range and median (bottom) for the first 50 positions with a clear separation between the non-random nucleotide positions of the tag (A), the quasi-random nucleotides of the tag (B) and the metagenomic sequence (C).

Simplified model showing how fragment-to-fragment concatenations can generate artificial concatenated sequences (left). Example of imperfect primer annealing that causes mismatch-induced mutations in the sequence reads (top).

Nucleotide frequency logos provided on the web interface that show the raw frequencies (top) and the filtered and corrected frequencies (bottom). The tag sequence can be identified more easily using filtered frequencies and allows more accurate predictions.