Abstract

High-throughput sequencing technologies are improving in quality, capacity, and costs, providing versatile applications in DNA research. For small genomes or fraction of larger genomes, DNA samples can be mixed and loaded together on the same sequencing track. This multiplexing approach relies on a specific DNA barcode that is attached to the sequencing adapter and accompanies every read. After sequencing, each sample read is identified on the basis of the respective barcode sequence.

Alterations of DNA barcodes during the experiment may lead to incorrect sample identification unless the error is corrected. This can be accomplished by implementing error correcting algorithms and codes. Two popular sets of error-correcting codes are Hamming codes and codes based on the Levenshtein distance.

Levenshtein-based codes operate only on words of known length. Since a DNA sequence with an embedded barcode is essentially one continuous long word, application of the classical Levenshtein algorithm is problematic. We demonstrate the decreased error correction capability of Levenshtein-based codes in a DNA context and suggest an adaptation of Levenshtein-based codes that is proven of efficiently correcting nucleotide errors in DNA sequences. In simulations we show the superior error correction capability of the new method compared to traditional Levenshtein and Hamming based codes in the presence of multiple errors.

The Illumina "Sequencing by Synthesis" platform shows a very large number of substitution errors as well as a very specific shift of the read that results in inserted and deleted bases at the 5'-end and the 3'-end. As a solution, we propose the *"Phaseshift distance"* that exclusively supports the correction of substitutions and phaseshifts.

We generated a large number of different sets of DNA barcodes using the Phaseshift distance and compared codes of different lengths and error correction capabilities. We found that codes based on the Phaseshift distance can correct a number of errors comparable to codes based on the Sequence-Levenshtein distance while offering the number of DNA barcodes comparable to Hamming codes.

In some cases, the position of the barcode and DNA context is not well defined. Many reads start inside the genomic insert so that adjacent primers might be missed. This is further complicated by coincidental similarities between barcode sequences and reference DNA. Therefore, a robust strategy is required in order to detect barcoded reads and avoid a large number of false positives or negatives.

The method presented in this thesis controls the tail area-based false discovery rate to distinguish between barcoded and unbarcoded reads. This method helps to establish the highest acceptable minimal distance between reads and barcode sequences. In a proof of concept experiment we estimated to correctly detect barcodes in 83% of the reads with a precision of 89%. Sensitivity improved to 99% at 99% precision when the adjacent primer sequence was incorporated in the analysis. The analysis was further improved using a paired end strategy. Following an analysis of the data for sequence variants induced in the *Atp1a1* gene of murine melanocytes, we found no evidence of cross-contamination of DNA material between samples.