CorQ users guide

General Description

CorQ is a set of perl programs that corrects errors in 454 pyrosequences by identifying and flagging poor quality insertions, deletions and substitutions within an alignment. Read coverage from both sequencing orientations is taken into consideration when correcting SNPs. Indels in homopolymer and non-homopolymer regions are checked for base quality and flagged if the algorithm identifies the indel as poor quality.

A nucleotide multiple sequence alignment is required for error correction. Mosaik is suggested for multiple sequence alignment.

Required External Software

Mosaik: Mosaik uses a Smith waterman based algorithm for alignment. Mosaik can be obtained at: http://bioinformatics.bc.edu/marthlab/Mosaik

Files needed for multiple sequence alignment: fasta and qual file with 454 reads, reference sequence.

Perl script to run Mosaik: align_with_Mosaik.pl
Required Input
-seq 454 sequence file
-qual Quality file
-out_seq Name for output files
-ref reference sequence for alignment
-rur File for unaligned sequences
-hs hash size for Mosaik (default is 10)
-minp minimum percentage of the read aligned (default is 0.7)
-mmp maximum mismatched percentage of read (default if 0.3)
-st sequencing technology(default is 454)

Usage
perl align_with_Mosaik.pl –seq <seq file> -qual <qual file> -out_seq <name of output> -ref <ref file> -rur <name of unaligned seq file> -hs <hash size> -minp <0 – 1> -mmp <0-1> -st <454, illumina, Sanger>

Output files
<Ace file>: Mosaik aligned sequences
<Mosaik log file>: File with Mosaik output and read alignment information

Script to convert Mosaik ace file into fasta format: convert_Mosaik_ace_to_fasta.pl
Required input
Input ace file
Name of output fasta file
Name of log file

Usage
perl convert_Mosaik_ace_to_fasta.pl <input ace file> <output fasta file> <output log file>

Output files
<Output.Fasta> file with aligned reads
<Output_short.fasta> Short read length files (Reads less than 100 bases)
<Output.stat> Output file with read coverage for each position in the alignment

Script to run base quality check: correct_poor_quality_indel_SNP.pl
Required Input
Ace file
Aligned fasta file
Fasta file
Quality file
Sample Name

Usage
perl correct_poor_quality_indel_SNP.pl <ace file> <aligned fasta file> <original fasta file> <original qual file> <sample name>

Optional parameters:
1. indel_num: this number will be considered to mark a region as part of a multiple indel. The default is 3.If three indels are found consecutively, then these will be considered as part of multiple indels and not corrected.
2. Fold_coverage: the default is 10. This value describes the read coverage difference between sequences from forward and reverse orientation. This value can be set to a number above 2.

Output files
<Log file>: File to store program correction information
<Flagged substitution file>: Stores positions of flagged substitution
<Flagged indel file>: stores positions of flagged indels
<Corrected fasta file>: Aligned file with corrected reads
<Annotation file>: File with annotations of changes made to each corrected sequence

Optional external software
AmpliconNoise: 454 Flowgram correction program
http://code.google.com/p/ampliconnoise/

Pyrobayes: Base quality recalibration program

http://bioinformatics.bc.edu/marthlab/PyroBayes

Test data
Fasta and quality file along with a reference sequence is provided to test the scripts in the following order:
1. align_with_Mosaik.pl
2. convert_Mosaik_ace_to_fasta.pl
3. correct_poor_quality_indel_SNP.pl