Karect v1.0
KAUST Assembly Read Error Correction Tool
(Instruction Manual)

Amin Allam
amin.allam@kaust.edu.sa

1 Description

Karect is a novel error correction tool based on multiple alignment. **Karect** supports substitution, insertion and deletion errors. It can handle non-uniform coverage as well as moderately covered areas of the sequenced genome. **Karect** supports *Illumina*, *454 FLX* and *Ion Torrent* sequencing machines. **Karect** includes an improved framework for evaluating the quality of error correction. Please cite this paper whenever appropriate:

> Allam,A. *et al* (2015) Karect: Accurate Correction of Substitution, Insertion and Deletion Errors for Next-generation Sequencing Data, *Bioinformatics*.

2 Main features

- **Karect** is easy to install; just type *make*.
- **Karect** supports substitution, insertion and deletion errors.
- **Karect** requires the least number of parameters.
- **Karect** can be configured to work in limited memory.
- **Karect** is accurate, efficient and reliable.

3 Error correction with Karect

3.1 Quick start

The following command can be used to run **Karect** on the input *fasta/fastq* file:

```
▷ ./karect -correct -threads=num_threads -matchtype=match_type \
      -celltype=cell_type -inputfile=input_file -resultdir=result_dir \
      -tempdir=temp_dir
```

where *match_type* is set to "hamming" for *Illumina* datasets, and "edit" for *454* and *Ion Torrent* datasets. *cell_type* is set to "diploid" or "haploid". *result_dir* and *temp_dir* must exist before running this command. You can specify several input files by repeating the option "-inputfile=*input_file*" for each file. Corrected file names will be located in *result_dir* and prefixed by "karect_".

3.2 Detailed description

We advice to stick with default values for all advanced options, unless you are very familiar with the tool. The general command is:

karect -correct [options]

Essential options	
-inputfile=f	▷ Specify an input *fasta/fastq* file. This option can be repeated for multiple files.
-celltype=x	▷ x=[haploid|diploid] Specify the cell type. Use *haploid* for bacteria and viruses. Used to better estimate coverage.
-matchtype=x	▷ x=[edit|hamming|insdel] Specify the matching type. *hamming* allows substitution errors only. *edit* allows insertions, deletion, and substitutions with equal costs. *insdel* is the same as *edit*, but the cost of substitutions is doubled. Use *hamming* for *Illumina* datasets, and *edit* for *454* and *Ion Torrent* datasets.

f=file, x=type

Basic options	
-inputdir=s	▷ Specify the input directory. Ignored if input file paths are complete [Default=.].
-resultdir=s	▷ Specify a directory to save result file(s). This directory must exist before running the command [Default=.].
-resultprefix=t	▷ Specify a prefix string to the result file(s) [Default=karect_].
-tempdir=s	▷ Specify a directory to save temporary output files. This directory must exist before running the command [Default=.].
-threads=i	▷ Specify the number of threads [Default=16].
-memory=d	▷ Specify an upper bound on the memory that can be used in gigabytes [Default=10240.0].

i=integer, d=decimal, s=directory, t=string

Advanced options	
-aggressive=d	▷ Specify the aggressiveness towards error correction. Increasing the value will increase *recall* and decrease *precision*. This constant is found in the equation of τ_3 in Section 2.6 of the paper [Default=0.42].
-numstages=i	▷ Specify the number of stages (1 or 2). 2 stages will cause a second stage of error correction (little more accurate but much slower) [Default=1].
-minoverlap=i	▷ Specify the minimum overlap size. This constant is found in the equation of τ_1 in Section 2.2 of the paper [Default=35].
-minoverlapper=d	▷ Specify the minimum overlap percentage. This constant is found in the equation of τ_1 in Section 2.2 of the paper [Default=0.20].
-minreadweigth=d	▷ Specify the minimum read weight [Default=0].
-errorrate=d	▷ Specify the first stage maximum allowed error rate. This constant is found in the equation of τ_2 in Section 2.2 of the paper [Default=0.25].
-errorratesec=d	▷ Specify the second stage maximum allowed error rate [Default=0.25].
-reserveval=d	▷ Specify the minimum reservation value. This constant is found in the equation of τ_3 in Section 2.6 of the paper [Default=100.0].
-estcov=x	▷ x=[yes|no] Estimate coverage and use it to adjust the minimum reservation value [Default=yes].
-usequal=x	▷ x=[yes|no] Use quality values of candidate reads [Default=yes].
-higherror	▷ Work in high error rate mode (error rate = 0.50).

i=integer, d=decimal, x=type

Advanced options (continued)	
-trimfact=d	▷ Specify the trimming factor. If trimming is enabled, read ending bases confirmed (by overlapping reads) with less than this value are trimmed [Default=2.5].
-reserveper=d	▷ Specify the minimum reservation percentage [Default=1.0].
-kmer=i	▷ Specify the minimum kmer size (will increase according to '-kmerfactor'). This value is equivalent to $l = k/3$ in Section 2.1 in the paper [Default=9].
-trim=x	▷ x=[yes|no] Allow/Disallow trimming. Do not allow trimming if evaluating results afterwards, or if you will pass results to an assembler which expects fixed read sizes [Default=no].

i=integer, d=decimal, x=type

More advanced options	
-maxlenmatches=i	▷ Specify the maximum number of expected alignment computations (millions) [Default=2000].
-maxkmerslots=i	▷ Specify the maximum number kmer slots to be used [Default=100,000].
-kmerfactor=i	▷ Specify the factor f such that $4^{kmersize} > total\_num\_kmers/f$ [Default=1000].
-maxkmerres=i	▷ Specify the maximum number kmer results to be used. This constant is found in the equation of m in Section 2.1 of the paper [Default=30].
-kmererrors=i	▷ Specify the maximum allowed kmer errors (0,1,2). This value is equivalent to d in Section 2.1 of the paper [Default=2].
-readsperstep=i	▷ Specify the maximum number of processed reads per step [Default=1000].
-fbs=i	▷ Specify file block size in megabytes [Default=10].
-cbs=i	▷ Specify cache block size in megabytes [Default=128].

i=integer

4 Evaluating error correction with Karect

4.1 Quick start

Error correction evaluation is done using the following commands:

```
▷ ./karect -align -threads=num_threads -matchtype=match_type \
      -inputfile=input_file -refgenomefile=ref_file -alignfile=align_file
▷ ./karect -eval -threads=num_threads -matchtype=match_type \
      -inputfile=input_file -resultfile=result_file \
      -refgenomefile=ref_file -alignfile=align_file -evalfile=eval_file
```

where *match_type* is set to "hamming" for *Illumina* datasets, and "edit" for *454* and *Ion Torrent* datasets. The first command aligns the input dataset to the reference genome, and produces the alignment file *align_file*. This command should be executed once per dataset. The second command evaluates the correction results *result_file* of a specific tool using alignment information of *align_file*. This command assumes that corrected reads in *result_file* are provided in the same order as the original reads *input_file*. The file *ref_file* is a *fasta* file containing the reference genome sequence(s). Currently, reference sequences up to 1 billion base-pairs in total are supported.

4.2 Detailed description for alignment

We advice to stick with default values for all advanced options, unless you are very familiar with the tool. The general command is:

▷ karect -align [options]

Essential options	
-matchtype=x	▷ x=[edit|hamming|insdel] Specify the matching type. *hamming* allows substitution errors only. *edit* allows insertions, deletion, and substitutions with equal costs. *insdel* is the same as *edit*, but the cost of substitutions is doubled.
-inputfile=f	▷ Specify an input *fasta*/*fastq* file. This option can be repeated for multiple files.
-refgenomefile=f	▷ Specify the *fasta* file containing the reference genome sequence(s) (to be aligned with). Currently, up to 1 billion bases in total are supported.
-alignfile=f	▷ Specify the output alignment file.

f=file, x=type

Basic options	
-inputdir=s	▷ Specify the input directory. Ignored if input file paths are complete [Default=.].
-threads=i	▷ Specify the number of threads [Default=16].

i=integer, s=directory

Advanced options	
-circular=i	▷ Specify the sequence size to be appended circularly (for circular genomes) [Default=0].
-accuracy=i	▷ Specify the alignment accuracy. The aligner is guaranteed to find the minimum distance alignments of all reads which can be mapped to the reference genome with less than $\lfloor readLen/(\lceil \log_4(2 \times refGenomeLen) \rceil - accuracy) \rfloor$ errors. Larger accuracy values can map more reads, but slower [Default=5].
-readsperstep=i	▷ Specify the maximum number of processed reads per step [Default=1000].

i=integer

4.3 Detailed description for evaluation

We advice to stick with default values for all advanced options, unless you are very familiar with the tool. The general command is:

▷ karect -eval [options]

Essential options	
-matchtype=x	▷ x=[edit|hamming|insdel] Specify the matching type. *hamming* allows substitution errors only. *edit* allows insertions, deletion, and substitutions with equal costs. *insdel* is the same as *edit*, but the cost of substitutions is doubled.
-inputfile=f	▷ Specify an input *fasta/fastq* file. This option can be repeated for multiple files.
-resultfile=f	▷ Specify a result *fasta/fastq* file (resulting from running "karect -correct", or any other correction tool). This option can be repeated for multiple files. The number and order of reads in *resultfile* must match the number and order of reads in *inputfile*.
-refgenomefile=f	▷ Specify the *fasta* file containing the reference genome sequence(s) (to be aligned with). Currently, up to 1 billion bases in total are supported.
-alignfile=f	▷ Specify the alignment file resulted from running "karect -align".
-evalfile=f	▷ Specify the output evaluation file.

f=file, x=type

Basic options	
-inputdir=s	▷ Specify the input files directory. Ignored if input file paths are complete [Default=.].
-resultdir=s	▷ Specify the result files directory. Ignored if result file paths are complete [Default=.].
-threads=i	▷ Specify the number of threads [Default=16].

i=integer, s=directory

Advanced options	
-circular=i	▷ Specify the sequence size to be appended circularly (for circular genomes) [Default=0].
-readsperstep=i	▷ Specify the maximum number of processed reads per step [Default=1000].

i=integer

5 Test data and complete running example

The following is a step-by-step description of a running example of correcting *Staphylococcus aureus Illumina* reads dataset:

1. Download the files *frag_1.fastq.gz* and *frag_2.fatstq.gz* (and *genome.fasta* if you need to evaluate results) from:
   http://gage.cbcb.umd.edu/data/Staphylococcus_aureus/Data.original/

2. Decompress *frag_1.fastq.gz* and *frag_2.fatstq.gz* by:

```
▷ gunzip frag_1.fastq.gz
▷ gunzip frag_2.fastq.gz
```

3. Use **Karect** to correct the read sequences (modify file paths if needed):

```
▷ ./karect -correct -threads=12 -matchtype=hamming \
      -celltype=haploid -inputfile=./frag_1.fastq -inputfile=./frag_2.fastq
```

   which produces the corrected read files: *karect_frag_1.fastq* and *karect_frag_2.fastq*

4. If you need to evaluate correction accuracy using the reference genome (*genome.fasta*):

   (a) Align original reads to the reference genome, to produce the file *align.txt*

```
▷ ./karect -align -threads=12 -matchtype=hamming \
      -inputfile=./frag_1.fastq -inputfile=./frag_2.fastq \
      -refgenomefile=./genome.fasta -alignfile=./align.txt
```

   (b) Evaluate the correction accuracy to produce the file *eval.txt*

```
▷ ./karect -eval -threads=12 -matchtype=hamming \
      -inputfile=./frag_1.fastq -inputfile=./frag_2.fastq \
      -resultfile=./karect_frag_1.fastq -resultfile=./karect_frag_2.fastq \
      -refgenomefile=./genome.fasta -alignfile=./align.txt \
      -evalfile=./eval.txt
```