GIbPSs 1.0 Manual Part 16: Documentation of createstdat_03.0.pl

5/22/2015 Andreas Hapke,
Institute of Anthropology, Johannes Gutenberg University Mainz, Germany, ahapke2@gmail.com

This program creates a simulated test dataset that you can analyze with **indloc** and further programs of GIbPSs as explained in the Tutorial in file 01_GIbPSs_1_0_overview.pdf.

Usage

Create a new directory, place a copy of the program in it, open a command prompt, cd to the directory and call **createstdat**:

```
createstdat_03.0.pl
```

or

```
perl createstdat_03.0.pl
```

The program creates a simulated test dataset and produces fastq files in a subdirectory `testdata`. It creates a second subdirectory `user_files` that contains files with suitable distance settings for **indloc** and **poploc** and a file for **indloc** about the fastq files.

The fastq files correspond to those that you would obtain with **fdm**, when you analyze paired sequencing data from a classical GBS library with the *rc-duplicates-strategy*. The simulated data have much less complex properties than real data but enable you to test many functions of GIbPSs quickly.

The simulated dataset comprises 1200 loci and 20 individuals. All loci have a length of 120 nucleotides and 2 to 6 different alleles, which differ by up to 6 SNPs. 200 loci contain indel variation: one allele has a deletion of 3 nucleotides at a randomly selected position. **createstdat** simulates individual genotypes and samples 40 read pairs per individual and locus. It samples reads from the two alleles of a locus at random. Forward- and reverse reads stem from one or the other DNA strand with a probability of 0.5. All reads have a length of 80 nucleotides and contain errors with a frequency of 0.001. The program constructs *rc-duplicates* and saves their forward reads to fastq-files.