Manual for fasta_stats_N50.pl

Minou Nowrousian

Feburary 2012

Summary:

The Perl program trim_fastq.pl calculates some basic statistics from a file with multiple nucleic acid sequences in fasta format (e.g. genomic contigs or scaffolds).

Usage: `fasta_stats_N50.pl file.fasta`

The following information is given in the output file (named file.fasta_stats.txt):

a) for the complete sequence
- number of sequences
- GC content in %
- total length of all sequences in bases
- N50 in bases
  The N50 value is an important quality metrics e.g. for genome assemblies. It is the value where 50 % of all bases in the assembly are in a contig of at least that size.
- number of gaps
- total length of all gaps in bases
- total gap length in %

b) for individual sequences
- fasta header
- length of sequence in bases
- GC content in %
- number of gaps
- length of gaps in bases
- length of gaps in %

Licensing information