**ART:** ART is a set of simulation tools to generate synthetic next-generation sequencing reads. ART supports simulation of single-end, paired-end/mate-pair reads of Illumina's Solexa, Roche's 454 and Applied Biosystems' SOLiD.

ART is implemented in C++ and Perl with optimized algorithms and is highly efficient in read simulation. ART inputs reads in the FASTA format and outputs read in the FASTQ format, and alignments in the ALN format.

ART simulates both single-end and paired-end sequencing reads of the three main commercial next-generation sequencing platforms: 454, Illumina and SOLiD. ART supports all three types of common sequencing errors: base substitutions, insertions, and deletions.

**Illumina read simulation:** This technology reads out one base at a time, the main error mode is substitution rather than insertion or deletion. The probability of a substitution error is determined by the base quality score associated with the called base. The distribution of base quality scores is position-dependent: the mean quality score decreases as a function of increasing base position.

ART simulates substitution errors according to the empirical, position-dependent distribution of base quality scores, measured in large training datasets. The base quality score does not directly provide information for INDEL errors, and ART simulates insertion and deletion errors directly from empirical distributions from our training data.

For paired-end simulation, ART uses two different quality score distributions and error rates for the first and second reads, each determined empirically.

**454 read simulation:** ART models the 454 sequencing error profile with homo-polymer lengthdependent over-call (insertion) and under-call (deletion) error distributions, and models base quality profiles as homopolymer-length dependent first order Markov chains.

ART uses an empirical distribution of 454 read lengths. ART generates 454 reads with built-in distributions derived for the 454 GS FLX sequencer model.

**SOLiD read simulation:** SOLiD base caller reports nucleotide transition color codes, rather than nucleotide sequences and accordingly, ART also generates nucleotide transition codes, or “color-space” reads.

For paired-end read simulations, a Gaussian distribution is used to model the distribution of DNA fragment sizes.

The built-in empirical error profiles of SOLiD reads were derived from the read data generated at Applied Biosystems.

**PERFORMANCE:** The test was performed on a desktop computer with Intel Xeon 2.93GHz CPU, running a Linux operating system.

|  |  |  |  |
| --- | --- | --- | --- |
| **PLATFORM** | **READ LENGTH** | **RUNNING TIME (SEC)** | **SPEED**  **(#READS/SEC)** |
|  |  | **SINGLE PAIRED** | **SINGLE PAIRED** |
| **454** | VARIED | 491 616 | 7049 10490 |
| **ILLUMINA** | 50 BP | 290 300 | 55997 54130 |
| **SOLiD** | 33 BP | 728 696 | 33798 33870 |

ThE procedure took less than 12 minutes (see Table 1), with Illumina reads being the fastest, and SOLiD reads the slowest.

**To Install in Unix shell:**

sudo apt install art-nextgen-simulation-tools

**To install in Conda workstation:**

Conda install -c bioconda art

**ART SIMULATION TOOLS COMMANDS:**

**art\_illumina** [options] **-sam** **-i** <seq\_ref\_file> **-l** <read\_length> **-f** <fold\_coverage> **-ss** <sequencing\_system> **-o** <outfile\_prefix>

**art\_illumina** [options] **-sam** **-i** <seq\_ref\_file> **-l** <read\_length> **-f** <fold\_coverage> **-o** <outfile\_prefix>

**art\_illumina** [options] **-sam** **-i** <seq\_ref\_file> **-l** <read\_length> **-c** <total\_num\_reads> **-o** <outfile\_prefix>

**art\_illumina** [options] **-sam** **-i** <seq\_ref\_file> **-l** <read\_length> **-f** <fold\_coverage> **-m** <mean\_fragsize> **-s** <std\_fragsize> **-o** <outfile\_prefix>

**art\_illumina** [options] **-sam** **-i** <seq\_ref\_file> **-l** <read\_length> **-c** <total\_num\_reads> **-m** <mean\_fragsize> **-s** <std\_fragsize> **-o** <outfile\_prefix>

**OPTIONS**

**-1** **--qprof1**

the first-read quality profile

**-2** **--qprof2**

the second-read quality profile

**-amp** **--amplicon** amplicon sequencing simulation

**-c** **--rcount**

total number of reads/read pairs to be generated [per amplicon if for amplicon simulation](not be used together with **-f**/--fcov)

**-d** **--id**

the prefix identification tag for read ID

**-ef** **--errfree**

indicate to generate the zero sequencing errors SAM file as well the regular one

NOTE: the reads in the zero-error SAM file have the same alignment positions as those in the regular SAM file, but have no sequencing errors

**-f** **--fcov**

the fold of read coverage to be simulated or number of reads/read pairs generated for each amplicon

**-h** **--help**

print out usage information

**-i** **--in**

the filename of input DNA/RNA reference

**-ir** **--insRate**

the first-read insertion rate (default: 0.00009)

**-ir2** **--insRate2** the second-read insertion rate (default: 0.00015)

**-dr** **--delRate**

the first-read deletion rate (default: 0.00011)

**-dr2** **--delRate2** the second-read deletion rate (default: 0.00023)

**-l** **--len**

the length of reads to be simulated

**-m** **--mflen**

the mean size of DNA/RNA fragments for paired-end simulations

**-mp** **--matepair** indicate a mate-pair read simulation

**-nf** **–maskN**

the cutoff frequency of 'N' in a window size of the read length for masking genomic regions

NOTE: default: '-nf 1' to mask all regions with 'N'. Use '-nf 0' to turn off masking

**-na** **--noALN**

do not output ALN alignment file

**-o** **--out**

the prefix of output filename

**-p** **--paired**

indicate a paired-end read simulation or to generate reads from both ends of amplicons

NOTE: art will automatically switch to a mate-pair simulation if the given mean fragment size >= 2000

**-q** **--quiet**

turn off end of run summary

**-qs** **--qShift**

the amount to shift every first-read quality score by

**-qs2** **--qShift2**

the amount to shift every second-read quality score by

NOTE: For **-qs**/-qs2 option, a positive number will shift up quality scores (the max is 93) that reduce substitution sequencing errors and a negative number will shift down quality scores that increase sequencing errors. If shifting scores by x, the error rate will be 1/(10^(x/10)) of the default profile.

**-rs** **--rndSeed**

the seed for random number generator (default: system time in second)

NOTE: using a fixed seed to generate two identical datasets from different runs

**-s** **--sdev**

the standard deviation of DNA/RNA fragment size for paired-end simulations.

**-sam** **--samout**

indicate to generate SAM alignment file

**-sp** **--sepProf**

indicate to use separate quality profiles for different bases (ATGC)

**-ss** **--seqSys**

The name of Illumina sequencing system of the built-in profile used for simulation

NOTE: sequencing system id names are:

GA1 - Genome Analyzer I, GA2 - Genome Analyzer II

HS10 - HiSeq 1000, HS20 - HiSeq 2000, HS25 - HiSeq 2500, MS - MiSeq

**-M** **--cigarM**

indicate to use CIGAR 'M' instead of '=/X' for alignment match/mismatch

**NOTES**

\* ART by default selects a built-in quality score profile according to the read length specified for the run.

\* For single-end simulation, ART requires input sequence file, outputfile prefix, read length, and read count/fold coverage.

\* For paired-end simulation (except for amplicon sequencing), ART also requires the parameter values of the mean and standard deviation of DNA/RNA fragment lengths

**EXAMPLES**

1) single-end read simulation

art\_illumina **-sam** **-i** reference.fa **-l** 150 **-ss** HS25 **-f** 10 **-o** single\_dat

2) paired-end read simulation

art\_illumina **-sam** **-i** reference.fa **-p** **-l** 150 **-ss** HS25 **-f** 20 **-m** 200 **-s** 10 **-o** paired\_dat

3) mate-pair read simulation

art\_illumina **-sam** **-i** reference.fa **-mp** **-l** 50 **-f** 20 **-m** 2500 **-s** 50 **-o** matepair\_dat

4) amplicon sequencing simulation with 5' end single-end reads

art\_illumina **-amp** **-sam** **-na** **-i** amp\_reference.fa **-l** 50 **-f** 10 **-o** amplicon\_5end\_dat

5) amplicon sequencing simulation with paired-end reads

art\_illumina **-amp** **-p** **-sam** **-na** **-i** amp\_reference.fa **-l** 50 **-f** 10 **-o** amplicon\_pair\_dat

6) amplicon sequencing simulation with matepair reads

art\_illumina **-amp** **-mp** **-sam** **-na** **-i** amp\_reference.fa **-l** 50 **-f** 10 **-o** amplicon\_mate\_dat

7) generate an extra SAM file with zero-sequencing errors for a paired-end read simulation

art\_illumina **-ef** **-i** reference.fa **-p** **-l** 50 **-f** 20 **-m** 200 **-s** 10 **-o** paired\_twosam\_dat

8) reduce the substitution error rate to one 10th of the default profile

art\_illumina **-i** reference.fa **-qs** 10 **-qs2** 10 **-l** 50 **-f** 10 **-p** **-m** 500 **-s** 10 **-sam** **-o** reduce\_error

9) turn off the masking of genomic regions with unknown nucleotides 'N'

art\_illumina **-nf** 0 **-sam** **-i** reference.fa **-p** **-l** 50 **-f** 20 **-m** 200 **-s** 10 **-o** paired\_nomask

10) masking genomic regions with >=5 'N's within the read length 50

art\_illumina **-nf** 5 **-sam** **-i** reference.fa **-p** **-l** 50 **-f** 20 **-m** 200 **-s** 10 **-o** paired\_maskN5