SPECTACLE – USER GUIDE

SPECTACLE: DATASETS

1. Illumina

ID	Reference		Links
	Species	G_L (Mbp)	
I1-10X	*R. sphaeroids*	4.6	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i1.04052015.tar.bz2
I1-20X			
I1-30X			
I1-40X			
I2-10X	*B. cereus ATCC 10987*	5.4	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i2.04052015.tar.bz2
I2-20X			
I2-30X			
I2-40X			
I3-10X	*O. sativa Chr. 5*	29.9	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i3.04052015.tar.bz2
I3-20X			
I3-30X			
I3-40X			
I4-10X	*Mouse Chr. Y*	88.1	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i4.04052015.tar.bz2
I4-20X			
I4-30X			
I4-40X			
I5-10X	*Human Chr. 1*	230.5	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i5.04052015.tar.bz2
I5-20X			
I5-30X			
I5-40X			
I6	*B. cereus ATCC 10987*	5.4	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/i5.04052015.tar.bz2

2. PacBio

ID	Reference		Links
	Species	G_L (Mbp)	
P1-10X	*E. coli*	4.6	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/p1.04052015.tar.bz2
P1-20X			
P1-30X			
P1-40X			

P2-10X P2-20X P2-30X P2-40X	*Human Chr19 10 Mbp*	10.0	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/p2.04052015.tar.bz2

3. ONT

ID	Reference		Links
	Species	G_L (Mbp)	
O1-10X O1-20X O1-30X O1-30x-EF	*E. coli*	4.6	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/o1.tar.bz2
O2-10X O2-20X O2-30X O2-30X-EF	*Saccharomyces cerevisiae W303*	7.5	http://spectacle.csl.illinois.edu/user/yunheo1/spectacle/data/o2.tar.bz2

SPECTACLE: SCRIPTS AND COMMANDS

1. Prerequisites

1.1. GCC

SPECTACLE was tested with gcc 4.9.2.

1.2 MPI package

SPECTACLE was tested with both OpenMPI 1.8.2 and MPICH 3.1.3 (GCC 4.7.1).

1.3. Perl modules

- Bio::DB::Sam (http://search.cpan.org/~lds/Bio-SamTools/lib/Bio/DB/Sam.pm)
- IO::Uncompress::Gunzip (http://search.cpan.org/~pmqs/IO-Compress-2.064/lib/IO/Uncompress/Gunzip.pm)

- Parallel::MPI::Simple (http://search.cpan.org/~ajgough/Parallel-MPI-Simple/Simple.pm)

  Installing this module may cause a problem. Makefile.PL executes "mpicc -help" to file compile options but the command may give nothing. In this case, Makefile.PL should be modified. See parallel-mpi-simple/Makefile.PL in the SPECTACLE directory.
- Sys::CPU (http://search.cpan.org/~mzsanford/Sys-CPU/CPU.pm)

1.4. SAMtools

In order to use the Bio::DB::Sam module, users need to compile SAMtools with the "-fPIC" option. The option should be added to CFLAGS in the Makefile.

1.5 SWIG

The location of the swig (http://www.swig.org) binary should be defined in $PATH. SPECTACLE was tested with SWIG 3.0.5.

2. Installing SPECTACLE

```
tar zxvf spectacle.vXpXX.tgz
cd vXpXX
Make
```

3. Usage

3.1. bam2location.common

Converts a BAM file that are generated using paired-end reads to an error location file.

Option	Description
bam <file>	Input BAM file. All reads that are not pairwise aligned should not be included in the file. Required.
errorfree	Generate error-free fasta files.
h	Prints a help message.
prefix <string>	Output file name prefix. Required.
q1 <file>	Input foward fastq file. Required.
q2 <file>	Input reverse fastq file. Required.
ref	Input reference sequence fasta file.
softclip	Convert softclipped bases into insertions.
thread	Number of threads for sorting. Default: number of cores.
tmp <dir>	Temporary directory name used internally for sorting. Default: system default directory.

3.2. bam2location.single.common

Converts a BAM file that are generated using single-end reads to an error location file.

Option	Description
bam <file>	Input BAM file. All reads that are not pairwise aligned should not be included in the file. Required.
errorfree	Generate a error-free fasta file.
h	Prints a help message.
prefix <string>	Output file name prefix. Required.
q <file>	Input fastq file. Required.
ref	Input reference sequence fasta file.
softclip	Convert softclipped bases into insertions.
thread	Number of threads for sorting. Default: number of cores.
tmp <dir>	Temporary directory name used internally for sorting. Default: system default directory.

3.3. compare-location-sam.dna

Compares the aligned locations in an input SAM file with those in an error location file.

3.3.1. Options

Option	Description
genome <1 or 2>	If reads came from two reference sequences, only reads that were sampled from the specified genome will be compared. If reads came from only one reference sequence, this value should be always 1. Required.
h	Prints a help message.
location <file>	Input error location file. Names of the Reads in this file should be matched with those in an input SAM file. Required.
noout	No output file. The program will print statistics to standard output.
prefix <string>	Output file name prefix. Required.
sam <file>	Input SAM file (pairwise aligned). Names of the reads in this file should be matched with those in an input error location file. Required.
strict	Uses the strict matching method. Aligned locations in an input error location file should be exactly matched with those in an input SAM file regardless of the existence of insertions or deletions.
thread	Number of threads for sorting. Default: number of cores.
t <dir>	Temporary directory name used internally for sorting. Default: system default directory.

3.3.2. Output reports

Item	Description
1.A	Number of read pairs that come from a specified genome. The genome can be specified using "-genome" option.
1.B	Number of read pairs that do not come from a specified genome. Zero for haploid genomes.
1.C	Number of read pairs that do not have corresponding information in the location file (i.e. "N/A" lines in the location file).
1.D	Number of read pairs that are aligned to a correct position.
1.E	Number of read pairs that are aligned to a wrong position.
1.F	Number of read pairs that are not aligned to any position.
2	Percentage of bases in the reference sequence that are covered by aligned read pairs. A read pair can be aligned to multiple positions. Read pairs that are aligned to wrong positions (i.e. 1. (E)) are also included.
3.1	Average number of reads that cover each base in the reference sequence. A read pair can be aligned to multiple positions.
3.2	Average number of reads that cover each base in the reference sequence. Only read pairs that are aligned to a correct position are considered.
4	Samtool binary used.

3.4. evaluate.dna

Evaluates DNA sequencing reads.

3.4.1. Options

Option	Description
bam1 <file>	Input BAM file generated using the flow in "Generating a BAM File for Coverage Analyses"
bam2 <file>	Input BAM file generated using the flow in "Generating a BAM File for Coverage Analyses"
candidate <number>	Max number of alignments that have the same highest score. If a read has alignments above this number, it read will not be evaluated. Default: 30,000
corfasta <file>	Input corrected single-end fasta file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfasta1 <file>	Input corrected forward read fasta file. The order of reads should be same as that of an original forward read file. Either corfasta, corfasta1 or corfastq1 should be used.
corfasta2 <file>	Input corrected reverse read fasta file. The order of reads should be same as that of an original reverse read file. Required if corfasta1 is used.

corfastq <file>	Input corrected single-end fastq file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfastq1 <file>	Input corrected forward read fastq file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfastq2 <file>	Input corrected reverse read fastq file. The order of reads should be same as that of an original reverse read file. Required if corfastq1 is used.
debug <prefix>	Generates debugging files with the specified prefix.
detailed <prefix>	Executes detailed accuracy analysis.
endgap	Penalize end gaps.
gext <number>	Gap extension penalty. Default: -1 (-1 for PacBio reads).
gopen <number>	Gap opening penalty. Default: -6 (-1 for PacBio reads).
h	Prints a help message.
location <file>	Input error location file. The name of reads in this file is not compared with that in input read files. However, the order of the reads should be same in both the files. Required.
-map <file>	Input original-corrected read mapping file. It has two columns; they show read names and the number of corresponding reads in the corrected read file.
match <number>	Match score. Default: 1 (1 for PacBio reads).
maxdepth <number>	Maximum range for reporting the detailed analysis results. Default: 50.
mmatch <number>	Mismatch penalty. Default: -4 (-1 for PacBio reads).
oneref	Load only one chromosome each time.
orgfasta <file>	Input original single-end fasta file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfasta1 <file>	Input original forward read fasta file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfasta2 <file>	Input original reverse read fasta file. Required if orgfasta1 is used.
orgfastq <file>	Input original single-end read fastq file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfastq1 <file>	Input original forward read fastq file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfastq2 <file>	Input original reverse read fastq file. Required if orgfastq1 is used.
outer <number>	The number of extra bases that are taken from a reference sequence to compensate insertions and deletions in the original read.
ref1 <file>	Input fasta file for the first reference sequence Ref 1. Required.
ref2 <file>	Input fasta file for the second reference sequence Ref 2. Required if input reads came from two reference sequences (i.e. Ref 1 and Ref 2).
tgs	Input reads are TGS reads
thread	Number of threads for sorting. Default: number of cores.

3.4.2. Output reports

The items in the output reports are explained below. Each base in corrected reads are categorized using triplets, each character of which should be either Y or N. The first position indicates whether the base in the original read is correct or not, respectively; the second position indicates whether the base has been modified by an error correction tool or not; and the third position indicates whether the base in the corrected read at that position is correct or not. All the bases should fall into one of five categories: NNN, NYN, NYY, YNY, and YYN; YYY, YNN, and NNY are logically impossible.

Supporting reads of base *b* in read *R* are the reads that overlap the corresponding base of *b* in a reference sequence (i.e. *R* and its supporting reads should be sampled from very close positions).

Items	Description
YYN	Number of bases 1) not erroneous in a precorrection read, 2) modified by an error correction tool, and 3) erroneous in a modified read.
YNY	Number of bases 1) not erroneous in a precorrection read, 2) not modified by an error correction tool, and 3) not erroneous in a modified read.
NYY	Number of bases 1) erroneous in a precorrection read, 2) modified by an error correction tool, and 3) not erroneous in a modified read.
NYY TRIM	NYY made by trimming.
NYN	Number of bases 1) erroneous in a precorrection read, 2) modified by an error correction tool, and 3) still erroneous in a modified read.
NNN	Number of bases 1) erroneous in a precorrection read, 2) not modified by an error correction tool, and 3) erroneous in a modified read.
From SUB to DEL	Number of bases 1) having substitution errors in precorrection reads and 2) changed to deletion errors in corrected reads.
Not evaluated	Number of errors in the reads that have more than 30,000 alignments with the same highest alignment score.
Sensitivity (recall)	NYY / (NYY + NYN + NNN)
Gain	(NYY - YYN - NYN) / (NYY + NYN + NNN)
Precision	NYY / (YYN + NYN)
Specificity	YNY / (YYN + YNY)

F-score	2NYY / (2NYY + YYN + 2NYN + NNN)
Position	Index: positions of errors in reads. As shown in **Error! Reference source not found.**, the numbers are the indices of corresponding bases in a reference sequence. Therefore, the number can be larger than read length when deletions exist in a read; Corrected bases: percentage of errors corrected.
(# of Correct Bases) - (# of Erroneous Bases)	Difference: number of supporting reads of a correct base - number of supporting reads of an erroneous base; Corrected bases: percentage of errors corrected.
# of Correct Bases	Coverage: number of supporting reads of a correct base, Corrected bases: percentage of errors corrected.

3.5. evaluation.rna

Evaluates RNA sequencing reads.

3.5.1. Options

Option	Description
candidate <number>	Max number of alignments that have the same highest score. If a read has alignments above this number, it read will not be evaluated. Default: 30,000
corfasta <file>	Input corrected single-end fasta file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfasta1 <file>	Input corrected forward read fasta file. The order of reads should be same as that of an original forward read file. Either corfasta, corfasta1 or corfastq1 should be used.
corfasta2 <file>	Input corrected reverse read fasta file. The order of reads should be same as that of an original reverse read file. Required if corfasta1 is used.
corfastq <file>	Input corrected single-end fastq file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfastq1 <file>	Input corrected forward read fastq file. The order of reads should be same as that of an original forward read file. Either corfasta, corfastq, corfasta1 or corfastq1 should be used.
corfastq2 <file>	Input corrected reverse read fastq file. The order of reads should be same as that of an original reverse read file. Required if corfastq1 is used.
debug <prefix>	Generates debugging files with the specified prefix.
endgap	Penalize end gaps.
gext <number>	Gap extension penalty. Default: -1 (-1 for TGS reads).

gopen <number>	Gap opening penalty. Default: -6 (-1 for TGS reads).
h	Prints a help message.
location <file>	Input error location file. The name of reads in this file is not compared with that in input corrected read files. However, the order of the reads should be same in both the files. Required.
match <number>	Match score. Default: 1 (1 for TGS reads).
mmatch <number>	Mismatch penalty. Default: -4 (-1 for TGS reads).
orgfasta <file>	Input original single-end fasta file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfasta1 <file>	Input original forward read fasta file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfasta2 <file>	Input original reverse read fasta file. Required if orgfasta1 is used.
orgfastq <file>	Input original single-end read fastq file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfastq1 <file>	Input original forward read fastq file. Either orgfasta, orgfastq, orgfasta1 or orgfastq1 should be used.
orgfastq2 <file>	Input original reverse read fastq file. Required if orgfastq1 is used.
pacbio	PacBio input reads.
ref <file>	Input error-free single-end fasta file. The order of reads should be same as that of an original forward read file. Either ref or reffasta1 should be used.
reffasta1 <file>	Input error-free forward fasta file. The order of reads should be same as that of an original forward read file. Either ref or reffasta1 should be used.
reffasta2 <file>	Input error-free reverse fasta file. The order of reads should be same as that of an original reverse read file. Required if reffasta1 is used.

3.5.2. Output reports

See 3.4.2.

3.6. fill-missed-reads.fasta.common

Fills missing reads in a corrected fasta file by copying them from an input precorrection read file and writes all the reads into an output fasta file.

Option	Description
h	Prints a help message.
corfasta <file>	Input corrected fasta file. Some reads in the file may be missing. Required.
orgfastq <file>	Input original fastq file. This file should contain all the reads missing in the corrected fasta file. Required.
outfasta <file>	Output fasta file. Required.

3.7. fill-missed-reads.fastq.common

Fills missing reads in a corrected fastq file by copying them from an input precorrection read file and writes all the reads into an output fastq file.

Option	Description
h	Prints a help message.
corfastq <file>	Input corrected fastq file. Some reads in the file may be missing. Required.
orgfastq <file>	Input original fastq file. This file should contain all the reads missing in the corrected fasta file. Required.
outfastq <file>	Output fastq file. Required.

3.8. generate-error-free-reads.dna

Generates error-free reads using an error location file. It checks where each read originates and takes that parts from a reference sequence.

Option	Description
h	Prints a help message.
location <file>	Input error location file. Required.
offset <33 or 64>	Quality score offset for an output file. If the value is 33 (64), all the quality socres are filled with "I" ("h"). Default: 33.
prefix <string>	Output file name prefix. Required.
ref1 <file>	Input fasta file for the first reference sequence Ref 1. Required.
ref2 <file>	Input fasta file for the second reference sequence Ref 2. Required if input reads came from two reference sequences (i.e. Ref 1 and Ref 2).

3.9. generate-map.fasta.common

Generates a map file that show how many reads the same name each read in the original file has.

Option	Description
h	Prints a help message.
corfasta <file>	Input corrected fasta file. Required.
orgfastq <file>	Input original fastq file. Required.
outmap <file>	Output map file. It has two columns; the first one is the read name and the second one is the number of occurrences of the read in corfasta. Required.

3.10. generate-map.fastq.common

Generates a map file that show how many reads the same name each read in the original file has.

Option	Description
h	Prints a help message.
corfastq <file>	Input corrected fastq file. Required.
orgfastq <file>	Input original fastq file. Required.
outmap <file>	Output map file. It has two columns; the first one is the read name and the second one is the number of occurrences of the read in corfastq. Required.

3.11. info2location.dna

Converts an info file generated by pIRS into an error location file.

Option	Description
h	Prints a help message.
info <file>	Input pIRS info file. Required.
location <file>	Input error location file. Required.
q1 <file>	Input foward fastq file. Required.
q2 <file>	Input reverse fastq file. Required.

3.12. modify-sam.common

Converts "=" or "X" in CIGAR strings in a input SAM file to "M"s.

Option	Description
h	Prints a help message.
in <file>	Input SAM file. Required.
out <file>	Output SAM file. Required.

3.13. remove-heterozygosity-from-location.common

Compare substitution errors in the input location file with heterozygous alleles in the input VCF file, and remove substitution errors that happened due to heterozygosities.

Option	Description
h	Prints a help message.
chr <string>	Chromosome that will be processed. Required.
location <file>	Input location file. Required.
vcf <file>	Input VCF file. Required
out <file>	Output location file. Required.

3.14. rename-fasta.common

Changes read names in a fasta file with those in another file.

Option	Description
h	Prints a help message.
infasta <file>	Input fasta file to be renamed. Required.
namefastq <file>	Input fastq file that has correct read names. The order of reads in this file should be same as those in the input infasta file. Required.
outfasta <file>	Output fasta file. Required.

3.15. rename-fasta.common.lsc

Remove postfix in read names in a LSC output file.

Option	Description
h	Prints a help message.
infasta <file>	Input fasta file to be renamed. Required.
namefastq <file>	Input fastq file that has correct read names. The order of reads in this file should be same as those in the input infasta file. Required.
outfasta <file>	Output fasta file. Required.

3.16. rename-fasta.common.pbcr

Remove postfix in read names in a PBcR output file.

Option	Description
h	Prints a help message.
infasta <file>	Input fasta file to be renamed. Required.
namefastq <file>	Input fastq file that has correct read names. The order of reads in this file should be same as those in the input infasta file. Required.
outfasta <file>	Output fasta file. Required.

3.17. rename-fasta.common.proovread

Remove postfix in read names in a Proovread output file.

Option	Description
h	Prints a help message.
infasta <file>	Input fasta file to be renamed. Required.
namefastq <file>	Input fastq file that has correct read names. The order of reads in this file should be same as those in the input infasta file. Required.
outfasta <file>	Output fasta file. Required.

3.18. rename-fastq.common

Changes read names in a fastq file with those in another file.

Option	Description
h	Prints a help message.
infastq <file>	Input fastq file to be renamed. Required.
namefastq <file>	Input fastq file that has correct read names. The order of reads in this file should be same as those in the infastq file. Required.
outfastq <file>	Output fastq file. Required.

3.19. reorder-fasta.common

Reorders reads in a fasta file to the same order in another file.

Option	Description
h	Prints a help message.
infasta <file>	Input fasta file to be reordered. Required.
orgfastq <file>	Input fastq file that has a correct read order. Reads in the input fasta file are reordered using the read order in this file. Required.
outfasta <file>	Output fasta file. Required.
tmp <dir>	Temporary directory name used internally for sorting. Default: system default directory.

3.20. reorder-fastq.common

Reorders reads in a fastq file to the same order in another file.

Option	Description
h	Prints a help message.
infastq <file>	Input fastq file to be reordered. Required.
orgfastq <file>	Input fastq file that has a correct read order. Reads in the input fasta file are reordered using the read order in this file. Required.
outfastq <file>	Output fastq file. Required.
tmp <dir>	Temporary directory name used internally for sorting. Default: system default directory.

3.21. similarity-bam.common

Calculate percent similarity of an input BAM file.

Option	Description
h	Prints a help message.
bam <file>	Input BAM file. Required.
fasta <file>	Input fasta file with which an input BAM file is generated. Either fasta or fastq should be used.
fastq <file>	Input fastq file with which an input BAM file is generated. Either fasta or fastq should be used.

3.22. simngs2location.rna

Finds errors that simNGS adds to error-free reads that are generated from BEERS.

Option	Description
errfastq <file>	Input fastq file from simNGS. The reads in the file have errors. Required.
errfreefasta <file>	Input fasta file from BEERS. The reads in the file have no error. Required.
gext <number>	Gap extension penalty. Default: -4.
gopen <number>	Gap opening penalty. Default: -11.
h	Prints a help message.
match <number>	Match score. Default: 4.
mmatch <number>	Mismatch penalty. Default: -1.
prefix <string>	Output file name prefix. Required.

3.23. split-ab.common

Splits reads according to the reference sequence from which they originate (i.e. Ref 1 or Ref 2).

Option	Description
1 <file>	Input forward read file. Required.
2 <file>	Input reverse read file. Required.
format <fasta | fastq>	Input read file format. Required.
h	Prints a help message.
location <file>	Input error location file. Required.
prefix <string>	Output file name prefix. Required.

3.24. unaligned2sam.common

Compares an input SAM file with input fastq files, finds unaligned reads, and adds the unaligned reads to the input SAM file.

Option	Description
fastq1	Input forward fastq file. Required.
fastq2	Input reverse fastq file. Required.
genome <1 or 2>	If reads came from two reference sequences, only reads that were sampled from the specified genome will be compared. Required.
h	Prints a help message.
location <file>	Input error location file. Required.
prefix <string>	Output file name prefix. Required.
ref <file>	Input reference fasta file. Required.
sam <file>	Input SAM file. Required.
samtools <file>	SAMtools binary. Required.
tmp <dir>	Temporary directory name used internally for sorting. Default: system default directory.