Coval User Manual

Ver. 1.4 25th November 2013

Contents

1. About Coval

Coval is a set of tools that filter and correct reads from SAM/BAM-formatted files and filter/call DNA variants (SNPs and short indels) from sam- or pileup-formatted files. Coval-Simulate creates a simulated genome that randomly introduces user-defined numbers of artificial SNPs and short indels. Coval-Refine and Coval-Call have been confirmed to work only for Illumina reads at present. The features of these tools are described as follows:

Coval-Refine, a tool to refine SAM/BAM alignment data (command: coval refine)

Coval-Refine, a SAM/BAM refinement tool, first tries to realign mismatch-containing reads around indels, and the realignment is fixed only when the number of mismatches contained in the realigned reads is decreased or unchanged, compared with the number of mismatches before realignment. After realignment, the program corrects potential sequencing errors in reads and removes reads that contain a larger number of mismatches than a user-specified mismatch number in a base quality-dependent manner. The program also removes reads containing at least three indels, one indel plus one soft-clipped end, or two soft-clipped ends. Although spliced alignment reads (represented with CIGAR containing 'N') can be also filtered, they are not used for both local realignment and error correction and indel-containing ones are predominantly filtered out. The program corrects errors in read that are judged from their base qualities and the allele frequency at the mismatch site before read filtering ('error correction mode'). In this mode, if there are

multiple non-reference alleles in a mismatch site, mismatch bases with lower allele frequencies at the site are corrected. The ‘error correction mode’ can be selected with ‘err_cor’ & ‘fpair’ or ‘ec_md’ options. The mismatch number and (quality-based) mismatch counting can be controlled by several options (-num, -fnum, -mrate, and -minq). This command also removes discordant paired-end reads for which either mate is unmapped or has a larger insert size than the average insert size plus 5-fold of the standard deviation, as default, although the removal of discordant paired-end reads may barely contribute to the improvement of SNP/indel calling accuracy. The average insert size and the standard deviation are automatically calculated from the first 5,000 concordantly aligned reads in a sorted SAM/BAM file. Implementation of the realignment function and the removal of discordant paired-end reads are selectable as options, as described in the ‘command line option’ section. When the realignment is prohibited, the tool further removes reads that contain mismatches in either of the terminal 2 bp regions of the reads, because many of these mismatch reads have resulted from misalignment around indels.

Input alignment files for the ‘refine’ command must be ‘sorted’ sam- or bam-formatted files, which can be generated by SAMtools (http://samtools.sourceforge.net/). Coval automatically judges whether the input file is sorted, and will abort if it is not sorted; however, when ‘realignment ’ and ‘read error correction’ is disabled (--dis_align is set but –err_cor not set), pre-sorting of SAM or BAM files is not needed. Coval does not restrict a read type (i.e., single-end or paired-end), read length, or alignment tools used for generation of the SAM/BAM alignment data, but may work well only for Illumina reads. Input SAM/BAM files need not have the MD tag because Coval-Refine does not see the MD tag for initial seeking of mismatches in

reads but outputs MD tags representing the altered mismatch information. The treatment of an alignment file containing 50 million reads with Coval-Refine should be accomplished within a few hours but it will take twice or more time when setting --msamp or --mfreq (>0) option in the error correction mode.

Coval-Call, a variant calling tool (command: coval call or coval call-sam)

The variant-calling tool implemented with the 'coval call' command filters and calls reliable DNA variants from a pileup variant file created with a 'pileup' (not mpileup) command in SAMtools or directly calls variants from a sorted SAM file. This tool extracts SNPs and short indels that support a user-specified minimum number of SNP/indel-supported reads, maximum number of covered reads, minimum allele frequency, and minimum base-call quality (only for SNPs). It then outputs two files that contain extracted SNPs and indels, respectively. The tool also filters out SNPs that exist within 3 bp of a called indel, and in the case of two neighboring indels that are 2 bp apart or closer, calls only the more reliable of the two. DNA variant calling from a pileup file containing 1.5 million variants would take about 1 min, and variant calling from a SAM file containing 50 millions reads would require about 4 h. The output from 'coval call' is slightly different from that from 'coval call-sam' because of a minor difference between their algorithms. We recommend using 'coval call', which performs the tasks more quickly.

Coval-Simulate, a simulation tool (command: coval simulate)

The simulation tool implemented with the 'coval simulate' command makes a simulated reference genome for use in

evaluating the accuracy of DNA variant calling. This tool randomly introduces user-specified rates of SNPs and 1 to 6 bp indels into a reference genome and outputs a FASTA file with the simulated genome and a text file containing the positions and bases of the introduced DNA variants. The most important feature of this tool is to introduce DNA variants at naturally occurring frequencies of SNP bases and indel lengths such that these variations in the simulated genome mimic real genome variations. It should take 4–5 h to create a simulated human genome containing 0.1% SNPs and 0.01% indels. It should be noted that there are in many cases differences in sequence (i.e., small or sometimes large numbers of DNA variants) between the public reference and the genome or set of RNAs that was sequenced. These endogenous DNA variants lead to a high background of false positives in the subsequent simulation analysis. To circumvent this problem, the observable endogenous variants should be eliminated from the reference genome by exchanging or modifying the reference bases with the bases of variants before generation of a simulated genome with Coval-Simulate.

2. Prerequisites

Coval can be implemented in Linux and MacOS systems.

(1) Perl 5.6 or later

All components of Coval are written in Perl. If your perl is not in the standard location (/usr/bin/perl), edit the first line of the Coval scripts or create a link of your perl executable to /usr/bin. Install perl modules (Getopt::Long, Pod::Usage, and File::Tee) if your system do not have these modules. If you need not output a log file, you may comment out lines 6, 124, and 125 in coval-refine-sam.pl.

(2) SAMtools-0.1.8 or before

SAMtools (http://sourceforge.net/projects/samtools/files/samtools/) is required for SAM/BAM format conversion, BAM sorting, and pileup variant calling. The PATH environmental variable must be set to your system. When using the pileup command of SAMtools, only samtools ver. 0.1.8 or before works properly, since ver. 0.1.9 or later seems to convert Phred base-call qualities from bam files wrongly. The pileup command has been removed from the latest version 0.1.17. Do not use SAMtools versions except for 0.1.8 or before when using the pileup command.

(3) Required input files

(a) coval refine

- ✓ Coordinate-sorted SAM or BAM alignment file
- ✓ Fasta file of reference sequence

(b) coval call

- ✓ Pileup file generated with samtools pileup

(c) coval call-sam

- ✓ Coordinate-sorted SAM file

(d) coval simulate

- ✓ Fasta file of reference sequence

3. Command line options

3.1 coval refine

filtering, error-correction, and local realignment of mismatch reads

<< Usage >>

coval refine [options] <input: sorted sam/bam file>

Input: a coordinate-sorted BAM or SAM file

	(Format conversion and bam sorting can be performed with SAMtools.) Caution! Read names and reference sequence names (the first and third columns) in the SAM file should not contain the characters '||' or '='.
Outout:	out_prefix.bam or out_prefix.sam

< Options >

--ref or -r <STR>	a reference FASTA file used for the alignment, which should be indexed with 'samtools faidx' (mandatory)
--pref or -p <STR>	prefix of output file (mandatory)
--sam or -s	outfile is sam [default: false]
--num or -n <INT>	the maximum number of allowable mismatches in a read [default: 2] (incompatible with --mrate) We have observed that the default value is optimal for references containing ~0.5% DNA variants and 75 bp Illumina reads. This value should be increased for more diverged references or longer reads. When using a value lower than the default (41) with the --minq option, the output would contain reads with a larger number of mismatches than the number specified with this option. Use --mrate option instead of --num when using long reads, such as 454 or Ion Torrent.
--mrate or -m <FLOAT>	the maximum rate of mismatches in a read [0..1.0] (incompatible with --num)

	This option would be useful when read lengths in an alignment are variable.
--fnum or -f <INT>	the maximum number of total mismatches contained in two paired reads [default: 1.7 * <INT> specified with --num] (incompatible with --mrate)
--fpair or -g	remove the other pair of a filtered (removed) paired-end read [default: false]
--qmap or -q <INT>	the minimum mapping quality. Reads with a mapping quality less than the specified number are filtered [default: 0] By default, read filtering according to mapping quality is not conducted. It is, however, recommended to set '-q 1' for calling heterozygous SNPs.
--qcall or -e <STR>	base call quality format of reads represented in sam/bam file; illumina (Phred+64) or sanger (Phred+33) [default: auto]
--minq or -k <INT>	the minimum base-call quality of mismatch base (counted as 'mismatch' only mismatched bases with the base quality lower than the specified value) [default: 41, which disables this filtering]
--avins or -a <STR>	An average read distance (a distance between the first reference positions mapped by two mates of paired-end reads) and its standard deviation,

	separated by a comma (e.g., 250,20) [default: auto, the distance and SD are automatically determined with a stat_insert.pl script.]
--insd or -i <INT>	a value M for setting the maximum read distance to be filtered: mean read distance + SD * M [default: 5] Read distance is the length between 5' termini of two paired-end reads, presented in SAM files. Both the read distance and its SD are automatically calculated by Coval-Refine and will be shown on the console.
--type or -t <FLOAT>	read type, paired-end (PE) or single-end (SE) [default: PE]
--reftype or –rt <COMPLETE or DRAFT>	forces --cdisc option when specifying DRAFT; DRAFT: draft-level of genomes or cDNAs/ESTs (e.g., genome assemblies), COMPLETE: finished genome sequences (e.g., human reference genome) This option is equal to specifying –c option and prevents the loss of reads aligned to terminal regions of a reference with a relatively short length. Do not specify this option if you want to remove discordantly aligned reads. [default: COMPLETE]
--cdisc or -c	include (not filter out) discordant paired-end reads, where either mate is unmapped or aligned to a different chromosome or the distance between

	the pair is highly deviated from an estimated distance [default: false] The use of this option may be useful for discovery for structural variants with other specific tools or for alignments to fragmented reference.
--unmap or -u	include (not filter out) unmapped reads [default: false]
--lins or -l	input reads contain mate-pair or paired-end reads with a long insert size [default: false]
--soap or -b <STR>	when alignment (sam format) was generated with SOAP aligner, use this option [default: false]
--err_cor or -x	correct potential sequencing errors in reads [default: false]
--qave or -xa <INT>	minimum mean base-call quality of mismatch bases covered at a potential non-reference allele in error correction mode [default: 10]
--mfreq or -xf <FLOAT>	minimum allele frequency at a potential non-reference allele in error correction mode [0..1.0] [default: 0] (Mismatch bases with lower quality and lower frequency than the values specified with -xa and -xf are corrected.)
--msamp or -ms <STR>	multiple (pooled) sample mode; calculate allele frequency for each sample based on RG tag. Read bases lower than the allele frequency specified with -xf are corrected [none, homo, or hetero] homo: for

	homozygous sample, -x -xf 0.8 is automatically set. hetero: for heterozygous sample, -x -xf 0.3 is set. [default: none]
--mallel or -ma	allow multiple non-reference alleles for each sample in multiple sample mode (e.g., alleles A, C, and T for samples-1, -2, and -3) [default: false]
--ec_md or -xm	set options suitable for 'error correction mode' (equal to '--err_cor --fpair') [default: false]
--talign_md or -tm	set options suitable for 'targeted' alignment (equal to '--fpair --fnum <INT specified with --num>') [default: false]
--dis_align or -d	disable realignment [default: false]
--help or -h	output help message

3.2 coval call and coval call-sam

filtering and calling of SNPs and short indels from pileup-formatted files or sorted sam files

<< Usage >>

coval call (or call-sam) [options] <input: pileup (or sorted sam file)>

Input (for 'call'):	a pileup file generated by the 'samtool pileup –vcf' command
Input (for 'call-sam'):	a sorted SAM file
Outout (for 'call'):	prefix-snp.pileup, prefix-indel.pileup
Outout (for 'call-sam'):	prefix-snp.txt, prefix-indel.txt (see Section 4 for format)

< Options >

--pref or -p <STR>	prefix of output files [default: out]
--num or -n <INT>	a minimum number of reads supporting a non-reference allele. An allele supported by fewer reads than the specified number is filtered. [default: 2]
--maxr or -m <INT>	the maximum read number covering a non-reference allele [default: 10000] In many cases, it is recommended to use 3-fold average read depth.
--freq or -f <FLOAT>	the minimum frequency of a non-reference allele [default: 0.8]
--tnum or -t <INT>	the minimum number of reads supporting a heterozygous non-reference allele [default: 3] This option is effective only when --freq is < 0.8.
--qual_base <INT>	the minimum allowable base-call quality of a non-reference base [default: 3]
--qual_ave <INT>	the minimum average base-call quality at a called site for an SNP [default: 20]
--calltype or -c <STR>	a quality format for base calling Illumina: phred+64 (illumina 1.3-1.5); Sanger: phred+33 (illumina 1.8) [default: auto]
--help or -h	output help message

3.3 coval simulate

creation of simulated reference genomes

<< Usage >>

coval simulate [options] -r <reference fasta file>

Output: prefix.fa (simulated reference FASTA)
prefix.snp (text file showing the positions and bases of introduced variants, see Section 4 for format)

< Options >

--ref or -r <STR>	a reference FASTA file (mandatory)
--pref or -p <STR>	prefix of output files [default: out]
--snp or -s <FLOAT>	the genomic mutation rate for SNPs [default: 0] (e.g., 0.001 for 0.1%)
--indel or -i <FLOAT>	the genomic mutation rate for indels [default: 0]
--help or –h	output help message

4. Output format

4.1 coval call-sam output

The coval call-sam command produces two output files, each with the output prefix name assigned by users, followed by '-snp.txt' or '-indel.txt'. These output files are different from pileup format files from 'coval call' or 'samtools pileup' and take the form of a text file containing nine tab-delimited fields with each line showing a filtered SNP or indel. The description of each field, which is common to both snp and indel files, is explained below.

Col	Field	Description
1	CHR	Chromosome name.
2	POS	Position. For an indel, position just before the indel.
3	REF	Reference base at POS. For a deletion, reference sequence corresponding to the deletion. For an

		insertion, '–' is indicated.
4	ALT	Non-reference base (allele) called at POS. For an insertion or deletion, the insertion sequence or '–' are indicated, respectively.
5	ALTN	Number of reads supporting ALT at POS.
6	RD	Number of reads covering POS.
7	ALTF	Frequency of ALT (i.e., ALT/RD).
8	MAPQ	Average mapping quality of reads covering POS.
9	QUAL	Average Phred quality of ALT.

4.2 coval simulate output

The 'coval simulate' command produces two output files, each with an output prefix name defined by users, followed by a '.fa' or '.snp' suffix. The former is a simulated reference FASTA file into which artificial SNPs and indels have been introduced. The latter is a text file, with each line showing the position and base of the introduced SNP and indel. The lines contain five tab-delimited fields, which are explained below.

Col	Field	Description
1	CHR	Chromosome name.
2	POS	Position. For an indel, position just before the indel.
3	VAR	Variant type, S: SNP, I: insertion, D: deletion.
4	REF	Reference base at POS. For a deletion, deleted reference sequence. For an insertion, '–' is indicated.
5	ALT	SNP base or insertion sequence introduced at POS.

5. Command examples

(1) < conversion of a sam alignment file to sorted bam file >

samtools view –uSt indexed_reference.fa.fai input.sam |
samtools sort – out_prefix

(2) < treatment of bam with 'coval refine' >

coval refine –r reference.fa –p output_prefix
input_sorted.bam

(3) < generate a pileup variant calling file >

samtools (e.g., ver. 0.1.8) pileup –vcf reference.fa
refined.bam > variant.pileup

(4) < filter variants from a pileup file >

(for read coverage of depth = 20x)

coval call variant.pileup –p out_prefix –m 60

or

< directly call variants from coval-treated sam >

coval call-sam refined.sam –p out_prefix –m 60

<< Recommended coval refine options for specific alignment data >>

(1) For a high diversity genome (e.g., containing 0.5–5% variants) or a long read length (e.g., > 100 bp)

coval refine –n (specify 3~5)

(2) For calling clustered variants (when N variants reside within a range of read length)

coval refine –n <N-1> –f <N*2>

(3) For 'targeted' alignment data (e.g., alignment of whole genome sequencing reads against a local chromosome fragment)

coval refine –g –f 2 (when using –n 2)

or

coval refine –tm

6. Performance test with sample dataset

A sample dataset for Coval performance test is available at http://sourceforge.net/projects/coval105/. The dataset contains a bam alignment file and its reference fasta file corresponding the chromosome 9 of a rice simulated genome, which was used in the Coval paper. Briefly, artificial SNPs and indels, corresponding 0.2 and 0.02% of the genome, (whose positional information can be obtained from the accompanied files, rice_simugenome_chr09.snp and rice_simugenome_chr09.indel) were introduced into the rice reference (IRGSP build 5) to generate a rice simulated genome. Experimentally obtained rice reads (6.3×10^7 75 bp paired-end reads) were aligned to the simulated genome, and the alignment data corresponding chromosome 9 were extracted. The included bam file has been sorted and PCR-duplicated reads have been removed. This bam file can be used for input file of Coval-Refine, followed by SNP/indel calling with Coval-Call or other variant callers. The accuracy of the called SNPs and indels can be determined by measuring the concordance of positions indicated in the included files, rice_simugenome_chr09.snp and rice_simugenome_chr09.indel, respectively.

7. Change logs

Ver.1.4:

(1) A problem associated with samtools in Coval-Refine that occurred on Macintosh machines was fixed.

(2) Reads aligned on a terminus of a reference chromosome

sometimes have wrong CIGAR strings with indels. The Coval-Refine was modified to change these indel-containing alignments to soft-clipped alignments (e.g., 80M2D20M -> 80M20S).

Ver.1.3:

(1) An option --reftype/-rt in Coval-Refine was newly added, which forces the --cdisc/-c option for a draft-level genome or a cDNA/EST set of reference and which prevents the loss of reads aligned to terminal regions of a reference with a relatively short length.

(2) A function to adjust the line length of a reference fasta file was added. This helps an automatic indexing of a reference file with 'samtools faidx' in Coval-Refine, which would fail to index fasta files with a long line length.

(3) A problem occurring when specifying the –mrate option was fixed.

8. Contact

Shunichi Kosugi　Kazusa DNA Research Institute
Email: skosugi@kazusa.or.jp