***Manual Reference Pages  -*samtools (1)**

**NAME**

samtools - Utilities for the Sequence Alignment/Map (SAM) format

bcftools - Utilities for the Binary Call Format (BCF) and VCF

**CONTENTS**

[Synopsis](http://samtools.sourceforge.net/samtools.shtml#1)  
[Description](http://samtools.sourceforge.net/samtools.shtml#2)  
[Samtools Commands And Options](http://samtools.sourceforge.net/samtools.shtml#3)  
[Bcftools Commands And Options](http://samtools.sourceforge.net/samtools.shtml#4)  
[Sam Format](http://samtools.sourceforge.net/samtools.shtml#5)  
[Vcf Format](http://samtools.sourceforge.net/samtools.shtml#6)  
[Examples](http://samtools.sourceforge.net/samtools.shtml#7)  
[Limitations](http://samtools.sourceforge.net/samtools.shtml#8)  
[Author](http://samtools.sourceforge.net/samtools.shtml#9)  
[See Also](http://samtools.sourceforge.net/samtools.shtml#10)

**SYNOPSIS**

samtools view -bt ref\_list.txt -o aln.bam aln.sam.gz

samtools sort aln.bam aln.sorted

samtools index aln.sorted.bam

samtools idxstats aln.sorted.bam

samtools view aln.sorted.bam chr2:20,100,000-20,200,000

samtools merge out.bam in1.bam in2.bam in3.bam

samtools faidx ref.fasta

samtools pileup -vcf ref.fasta aln.sorted.bam

samtools mpileup -C50 -gf ref.fasta -r chr3:1,000-2,000 in1.bam in2.bam

samtools tview aln.sorted.bam ref.fasta

bcftools index in.bcf

bcftools view in.bcf chr2:100-200 > out.vcf

bcftools view -vc in.bcf > out.vcf 2> out.afs

**DESCRIPTION**

Samtools is a set of utilities that manipulate alignments in the BAM format. It imports from and exports to the SAM (Sequence Alignment/Map) format, does sorting, merging and indexing, and allows to retrieve reads in any regions swiftly.

Samtools is designed to work on a stream. It regards an input file ‘-’ as the standard input (stdin) and an output file ‘-’ as the standard output (stdout). Several commands can thus be combined with Unix pipes. Samtools always output warning and error messages to the standard error output (stderr).

Samtools is also able to open a BAM (not SAM) file on a remote FTP or HTTP server if the BAM file name starts with ‘ftp://’ or ‘http://’. Samtools checks the current working directory for the index file and will download the index upon absence. Samtools does not retrieve the entire alignment file unless it is asked to do so.

**SAMTOOLS COMMANDS AND OPTIONS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **view** | | samtools view [-bchuHS] [-t in.refList] [-o output] [-f reqFlag] [-F skipFlag] [-q minMapQ] [-l library] [-r readGroup] [-R rgFile] <in.bam>|<in.sam> [region1 [...]]  Extract/print all or sub alignments in SAM or BAM format. If no region is specified, all the alignments will be printed; otherwise only alignments overlapping the specified regions will be output. An alignment may be given multiple times if it is overlapping several regions. A region can be presented, for example, in the following format: ‘chr2’ (the whole chr2), ‘chr2:1000000’ (region starting from 1,000,000bp) or ‘chr2:1,000,000-2,000,000’ (region between 1,000,000 and 2,000,000bp including the end points). The coordinate is 1-based.  **OPTIONS:**   |  | | --- | |  | |  | | **-b** | Output in the BAM format. | | **-f** *INT* | Only output alignments with all bits in INT present in the FLAG field. INT can be in hex in the format of /^0x[0-9A-F]+/ [0] | | **-F** *INT* | Skip alignments with bits present in INT [0] | | **-h** | Include the header in the output. | | **-H** | Output the header only. | | **-l** *STR* | Only output reads in library STR [null] | | **-o** *FILE* | Output file [stdout] | | **-q** *INT* | Skip alignments with MAPQ smaller than INT [0] | | **-r** *STR* | Only output reads in read group STR [null] | | **-R** *FILE* | Output reads in read groups listed in *FILE* [null] | | **-S** | Input is in SAM. If @SQ header lines are absent, the **‘-t’** option is required. | | **-c** | Instead of printing the alignments, only count them and print the total number. All filter options, such as **‘-f’,** **‘-F’** and **‘-q’** , are taken into account. | | **-t** *FILE* | This file is TAB-delimited. Each line must contain the reference name and the length of the reference, one line for each distinct reference; additional fields are ignored. This file also defines the order of the reference sequences in sorting. If you run ‘samtools faidx <ref.fa>’, the resultant index file *<ref.fa>.fai* can be used as this *<in.ref\_list>* file. | | **-u** | Output uncompressed BAM. This option saves time spent on compression/decomprssion and is thus preferred when the output is piped to another samtools command. | |  |  | |
| **tview** | | samtools tview <in.sorted.bam> [ref.fasta]  Text alignment viewer (based on the ncurses library). In the viewer, press ‘?’ for help and press ‘g’ to check the alignment start from a region in the format like ‘chr10:10,000,000’ or ‘=10,000,000’ when viewing the same reference sequence. |
| **mpileup** | | **samtools mpileup** [**-EBug**] [**-C** *capQcoef*] [**-r** *reg*] [**-f** *in.fa*] [**-l** *list*] [**-M** *capMapQ*] [**-Q** *minBaseQ*] [**-q** *minMapQ*] *in.bam* [*in2.bam* [*...*]]  Generate BCF or pileup for one or multiple BAM files. Alignment records are grouped by sample identifiers in @RG header lines. If sample identifiers are absent, each input file is regarded as one sample.  In the pileup format (without **-u**or**-g**), each line represents a genomic position, consisting of chromosome name, coordinate, reference base, read bases, read qualities and alignment mapping qualities. Information on match, mismatch, indel, strand, mapping quality and start and end of a read are all encoded at the read base column. At this column, a dot stands for a match to the reference base on the forward strand, a comma for a match on the reverse strand, a ’>’ or ’<’ for a reference skip, ‘ACGTN’ for a mismatch on the forward strand and ‘acgtn’ for a mismatch on the reverse strand. A pattern ‘\+[0-9]+[ACGTNacgtn]+’ indicates there is an insertion between this reference position and the next reference position. The length of the insertion is given by the integer in the pattern, followed by the inserted sequence. Similarly, a pattern ‘-[0-9]+[ACGTNacgtn]+’ represents a deletion from the reference. The deleted bases will be presented as ‘\*’ in the following lines. Also at the read base column, a symbol ‘^’ marks the start of a read. The ASCII of the character following ‘^’ minus 33 gives the mapping quality. A symbol ‘$’ marks the end of a read segment.  **Input Options:**   |  | | --- | |  | |  | | **-6** | Assume the quality is in the Illumina 1.3+ encoding. **-A** Do not skip anomalous read pairs in variant calling. | | **-B** | Disable probabilistic realignment for the computation of base alignment quality (BAQ). BAQ is the Phred-scaled probability of a read base being misaligned. Applying this option greatly helps to reduce false SNPs caused by misalignments. | | **-b** *FILE* | List of input BAM files, one file per line [null] | | **-C** *INT* | Coefficient for downgrading mapping quality for reads containing excessive mismatches. Given a read with a phred-scaled probability q of being generated from the mapped position, the new mapping quality is about sqrt((INT-q)/INT)\*INT. A zero value disables this functionality; if enabled, the recommended value for BWA is 50. [0] | | **-d** *INT* | At a position, read maximally *INT* reads per input BAM. [250] | | **-E** | Extended BAQ computation. This option helps sensitivity especially for MNPs, but may hurt specificity a little bit. | | **-f** *FILE* | The **faidx**-indexed reference file in the FASTA format. The file can be optionally compressed by **razip**. [null] | | **-l** *FILE* | BED or position list file containing a list of regions or sites where pileup or BCF should be generated [null] | | **-q** *INT* | Minimum mapping quality for an alignment to be used [0] | | **-Q** *INT* | Minimum base quality for a base to be considered [13] | | **-r** *STR* | Only generate pileup in region *STR* [all sites] | | **Output Options:** | | |  |  | | **-D** | Output per-sample read depth | | **-g** | Compute genotype likelihoods and output them in the binary call format (BCF). | | **-S** | Output per-sample Phred-scaled strand bias P-value | | **-u** | Similar to **-g** except that the output is uncompressed BCF, which is preferred for piping. | | **Options for Genotype Likelihood Computation (for -g or -u):** | | |  |  | | **-e** *INT* | Phred-scaled gap extension sequencing error probability. Reducing *INT* leads to longer indels. [20] | | **-h** *INT* | Coefficient for modeling homopolymer errors. Given an *l*-long homopolymer run, the sequencing error of an indel of size *s* is modeled as *INT*\**s*/*l*. [100] | | **-I** | Do not perform INDEL calling | | **-L** *INT* | Skip INDEL calling if the average per-sample depth is above *INT*. [250] | | **-o** *INT* | Phred-scaled gap open sequencing error probability. Reducing *INT* leads to more indel calls. [40] | | **-P** *STR* | Comma dilimited list of platforms (determined by **@RG-PL**) from which indel candidates are obtained. It is recommended to collect indel candidates from sequencing technologies that have low indel error rate such as ILLUMINA. [all] | |  |  | |
| **reheader** | | samtools reheader <in.header.sam> <in.bam>  Replace the header in *in.bam* with the header in *in.header.sam.* This command is much faster than replacing the header with a BAM->SAM->BAM conversion. |
| **cat** | | samtools cat [-h header.sam] [-o out.bam] <in1.bam> <in2.bam> [ ... ]  Concatenate BAMs. The sequence dictionary of each input BAM must be identical, although this command does not check this. This command uses a similar trick to **reheader** which enables fast BAM concatenation. |
| **sort** | | samtools sort [-no] [-m maxMem] <in.bam> <out.prefix>  Sort alignments by leftmost coordinates. File *<out.prefix>.bam* will be created. This command may also create temporary files *<out.prefix>.%d.bam* when the whole alignment cannot be fitted into memory (controlled by option -m).  **OPTIONS:**   |  | | --- | |  | |  | | **-o** | Output the final alignment to the standard output. | | **-n** | Sort by read names rather than by chromosomal coordinates | | **-m** *INT* | Approximately the maximum required memory. [500000000] | |  |  | |
| **merge** | | samtools merge [-nur1f] [-h inh.sam] [-R reg] <out.bam> <in1.bam> <in2.bam> [...]  Merge multiple sorted alignments. The header reference lists of all the input BAM files, and the @SQ headers of *inh.sam*, if any, must all refer to the same set of reference sequences. The header reference list and (unless overridden by **-h**) ‘@’ headers of *in1.bam* will be copied to *out.bam*, and the headers of other files will be ignored.  **OPTIONS:**   |  | | --- | |  | |  | | **-1** | Use zlib compression level 1 to comrpess the output | | **-f** | Force to overwrite the output file if present. | | **-h** *FILE* | Use the lines of *FILE* as ‘@’ headers to be copied to *out.bam*, replacing any header lines that would otherwise be copied from *in1.bam*. (*FILE* is actually in SAM format, though any alignment records it may contain are ignored.) | | **-n** | The input alignments are sorted by read names rather than by chromosomal coordinates | | **-R** *STR* | Merge files in the specified region indicated by *STR* [null] | | **-r** | Attach an RG tag to each alignment. The tag value is inferred from file names. | | **-u** | Uncompressed BAM output | |  |  | |
| **index** | | samtools index <aln.bam>  Index sorted alignment for fast random access. Index file *<aln.bam>.bai* will be created. |
| **idxstats** | | samtools idxstats <aln.bam>  Retrieve and print stats in the index file. The output is TAB delimited with each line consisting of reference sequence name, sequence length, # mapped reads and # unmapped reads. |
| **faidx** | | samtools faidx <ref.fasta> [region1 [...]]  Index reference sequence in the FASTA format or extract subsequence from indexed reference sequence. If no region is specified, **faidx** will index the file and create *<ref.fasta>.fai* on the disk. If regions are speficified, the subsequences will be retrieved and printed to stdout in the FASTA format. The input file can be compressed in the **RAZF** format. |
| **fixmate** | | samtools fixmate <in.nameSrt.bam> <out.bam>  Fill in mate coordinates, ISIZE and mate related flags from a name-sorted alignment. |
| **rmdup** | | samtools rmdup [-sS] <input.srt.bam> <out.bam>  Remove potential PCR duplicates: if multiple read pairs have identical external coordinates, only retain the pair with highest mapping quality. In the paired-end mode, this command **ONLY** works with FR orientation and requires ISIZE is correctly set. It does not work for unpaired reads (e.g. two ends mapped to different chromosomes or orphan reads).  **OPTIONS:**   |  | | --- | |  | |  | | **-s** | Remove duplicate for single-end reads. By default, the command works for paired-end reads only. | | **-S** | Treat paired-end reads and single-end reads. | |  |  | |
| **calmd** | | samtools calmd [-EeubSr] [-C capQcoef] <aln.bam> <ref.fasta>  Generate the MD tag. If the MD tag is already present, this command will give a warning if the MD tag generated is different from the existing tag. Output SAM by default.  **OPTIONS:**   |  | | --- | |  | |  | | **-A** | When used jointly with **-r** this option overwrites the original base quality. | | **-e** | Convert a the read base to = if it is identical to the aligned reference base. Indel caller does not support the = bases at the moment. | | **-u** | Output uncompressed BAM | | **-b** | Output compressed BAM | | **-S** | The input is SAM with header lines | | **-C** *INT* | Coefficient to cap mapping quality of poorly mapped reads. See the **pileup** command for details. [0] | | **-r** | Compute the BQ tag (without -A) or cap base quality by BAQ (with -A). | | **-E** | Extended BAQ calculation. This option trades specificity for sensitivity, though the effect is minor. | |  |  | |
| **targetcut** | | samtools targetcut [-Q minBaseQ] [-i inPenalty] [-0 em0] [-1 em1] [-2 em2] [-f ref] <in.bam>  This command identifies target regions by examining the continuity of read depth, computes haploid consensus sequences of targets and outputs a SAM with each sequence corresponding to a target. When option **-f** is in use, BAQ will be applied. This command is **only** designed for cutting fosmid clones from fosmid pool sequencing [Ref. Kitzman et al. (2010)]. |
|  | |  |
| **phase** | samtools phase [-AF] [-k len] [-b prefix] [-q minLOD] [-Q minBaseQ] <in.bam>  Call and phase heterozygous SNPs. **OPTIONS:**   |  | | --- | |  | |  | | **-A** | Drop reads with ambiguous phase. | | **-b** *STR* | Prefix of BAM output. When this option is in use, phase-0 reads will be saved in file **STR**.0.bam and phase-1 reads in **STR**.1.bam. Phase unknown reads will be randomly allocated to one of the two files. Chimeric reads with switch errors will be saved in **STR**.chimeric.bam. [null] | | **-F** | Do not attempt to fix chimeric reads. | | **-k** *INT* | Maximum length for local phasing. [13] | | **-q** *INT* | Minimum Phred-scaled LOD to call a heterozygote. [40] | | **-Q** *INT* | Minimum base quality to be used in het calling. [13] | |  |  | | | |
|  |  | | |

**BCFTOOLS COMMANDS AND OPTIONS**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **view** | | **bcftools view** [**-AbFGNQSucgv**] [**-D** *seqDict*] [**-l** *listLoci*] [**-s** *listSample*] [**-i** *gapSNPratio*] [**-t** *mutRate*] [**-p** *varThres*] [**-P** *prior*] [**-1** *nGroup1*] [**-d** *minFrac*] [**-U** *nPerm*] [**-X** *permThres*] [**-T** *trioType*] *in.bcf* [*region*]  Convert between BCF and VCF, call variant candidates and estimate allele frequencies.   |  |  | | --- | --- | | **Input/Output Options:**  **-A** | | |  | Retain all possible alternate alleles at variant sites. By default, the view command discards unlikely alleles. | |  |  |  |  |  | | --- | --- | | **-b** | Output in the BCF format. The default is VCF. | | **-D** *FILE* | Sequence dictionary (list of chromosome names) for VCF->BCF conversion [null] | | **-F** | Indicate PL is generated by r921 or before (ordering is different). | | **-G** | Suppress all individual genotype information. | | **-l** *FILE* | List of sites at which information are outputted [all sites] | | **-N** | Skip sites where the REF field is not A/C/G/T | | **-Q** | Output the QCALL likelihood format | | **-s** *FILE* | List of samples to use. The first column in the input gives the sample names and the second gives the ploidy, which can only be 1 or 2. When the 2nd column is absent, the sample ploidy is assumed to be 2. In the output, the ordering of samples will be identical to the one in *FILE*. [null] | | **-S** | The input is VCF instead of BCF. | | **-u** | Uncompressed BCF output (force -b). | | **Consensus/Variant Calling Options:**  **-c** | | |  | Call variants using Bayesian inference. This option automatically invokes option **-e**. | | **-d** *FLOAT* | When **-v** is in use, skip loci where the fraction of samples covered by reads is below FLOAT. [0] | | **-e** | Perform max-likelihood inference only, including estimating the site allele frequency, testing Hardy-Weinberg equlibrium and testing associations with LRT. | | **-g** | Call per-sample genotypes at variant sites (force -c) | | **-i** *FLOAT* | Ratio of INDEL-to-SNP mutation rate [0.15] | | **-p** *FLOAT* | A site is considered to be a variant if P(ref|D)<FLOAT [0.5] | | **-P** *STR* | Prior or initial allele frequency spectrum. If STR can be *full*, *cond2*, *flat* or the file consisting of error output from a previous variant calling run. | | **-t** *FLOAT* | Scaled muttion rate for variant calling [0.001] | | **-T** *STR* | Enable pair/trio calling. For trio calling, option **-s** is usually needed to be applied to configure the trio members and their ordering. In the file supplied to the option **-s**, the first sample must be the child, the second the father and the third the mother. The valid values of *STR* are ‘pair’, ‘trioauto’, ‘trioxd’ and ‘trioxs’, where ‘pair’ calls differences between two input samples, and ‘trioxd’ (‘trioxs’) specifies that the input is from the X chromosome non-PAR regions and the child is a female (male). [null] | | **-v** | Output variant sites only (force -c) | | **Contrast Calling and Association Test Options:**  **-1** *INT* | | |  | Number of group-1 samples. This option is used for dividing the samples into two groups for contrast SNP calling or association test. When this option is in use, the following VCF INFO will be outputted: PC2, PCHI2 and QCHI2. [0] | | **-U** *INT* | Number of permutations for association test (effective only with **-1**) [0] | | **-X** *FLOAT* | Only perform permutations for P(chi^2)<FLOAT (effective only with **-U**) [0.01] | |  |  | |
| **index** | | **bcftools index** *in.bcf*  Index sorted BCF for random access. |
|  | |  |
| **cat** | **bcftools cat** *in1.bcf* [*"in2.bcf "*[*...*"]]]"  Concatenate BCF files. The input files are required to be sorted and have identical samples appearing in the same order. | | |
|  |  | | |

**SAM FORMAT**

Sequence Alignment/Map (SAM) format is TAB-delimited. Apart from the header lines, which are started with the ‘@’ symbol, each alignment line consists of:

|  |  |  |
| --- | --- | --- |
| **Col** | **Field** | **Description** |
|  |  |  |
|  |  |  |
| 1 | QNAME | Query template/pair NAME |
| 2 | FLAG | bitwise FLAG |
| 3 | RNAME | Reference sequence NAME |
| 4 | POS | 1-based leftmost POSition/coordinate of clipped sequence |
| 5 | MAPQ | MAPping Quality (Phred-scaled) |
| 6 | CIAGR | extended CIGAR string |
| 7 | MRNM | Mate Reference sequence NaMe (‘=’ if same as RNAME) |
| 8 | MPOS | 1-based Mate POSistion |
| 9 | TLEN | inferred Template LENgth (insert size) |
| 10 | SEQ | query SEQuence on the same strand as the reference |
| 11 | QUAL | query QUALity (ASCII-33 gives the Phred base quality) |
| 12+ | OPT | variable OPTional fields in the format TAG:VTYPE:VALUE |

Each bit in the FLAG field is defined as:

|  |  |  |
| --- | --- | --- |
| **Flag** | **Chr** | **Description** |
|  |  |  |
|  |  |  |
| 0x0001 | p | the read is paired in sequencing |
| 0x0002 | P | the read is mapped in a proper pair |
| 0x0004 | u | the query sequence itself is unmapped |
| 0x0008 | U | the mate is unmapped |
| 0x0010 | r | strand of the query (1 for reverse) |
| 0x0020 | R | strand of the mate |
| 0x0040 | 1 | the read is the first read in a pair |
| 0x0080 | 2 | the read is the second read in a pair |
| 0x0100 | s | the alignment is not primary |
| 0x0200 | f | the read fails platform/vendor quality checks |
| 0x0400 | d | the read is either a PCR or an optical duplicate |

where the second column gives the string representation of the FLAG field.

**VCF FORMAT**

The Variant Call Format (VCF) is a TAB-delimited format with each data line consists of the following fields:

|  |  |  |
| --- | --- | --- |
| **Col** | **Field** | **Description** |
|  |  |  |
|  |  |  |
| 1 | CHROM | CHROMosome name |
| 2 | POS | the left-most POSition of the variant |
| 3 | ID | unique variant IDentifier |
| 4 | REF | the REFerence allele |
| 5 | ALT | the ALTernate allele(s), separated by comma |
| 6 | QUAL | variant/reference QUALity |
| 7 | FILTER | FILTers applied |
| 8 | INFO | INFOrmation related to the variant, separated by semi-colon |
| 9 | FORMAT | FORMAT of the genotype fields, separated by colon (optional) |
| 10+ | SAMPLE | SAMPLE genotypes and per-sample information (optional) |

The following table gives the **INFO** tags used by samtools and bcftools.

|  |  |  |
| --- | --- | --- |
| **Tag** | **Format** | **Description** |
|  |  |  |
|  |  |  |
| AF1 | double | Max-likelihood estimate of the site allele frequency (AF) of the first ALT allele |
| DP | int | Raw read depth (without quality filtering) |
| DP4 | int[4] | # high-quality reference forward bases, ref reverse, alternate for and alt rev bases |
| FQ | int | Consensus quality. Positive: sample genotypes different; negative: otherwise |
| MQ | int | Root-Mean-Square mapping quality of covering reads |
| PC2 | int[2] | Phred probability of AF in group1 samples being larger (,smaller) than in group2 |
| PCHI2 | double | Posterior weighted chi^2 P-value between group1 and group2 samples |
| PV4 | double[4] | P-value for strand bias, baseQ bias, mapQ bias and tail distance bias |
| QCHI2 | int | Phred-scaled PCHI2 |
| RP | int | # permutations yielding a smaller PCHI2 |
| CLR | int | Phred log ratio of genotype likelihoods with and without the trio/pair constraint |
| UGT | string | Most probable genotype configuration without the trio constraint |
| CGT | string | Most probable configuration with the trio constraint |

**EXAMPLES**

|  |  |
| --- | --- |
| o | Import SAM to BAM when **@SQ** lines are present in the header:  samtools view -bS aln.sam > aln.bam  If **@SQ** lines are absent:  samtools faidx ref.fa  samtools view -bt ref.fa.fai aln.sam > aln.bam  where *ref.fa.fai* is generated automatically by the **faidx** command. |
| o | Attach the **RG** tag while merging sorted alignments:  perl -e ’print "@RG\tID:ga\tSM:hs\tLB:ga\tPL:Illumina\n@RG\tID:454\tSM:hs\tLB:454\tPL:454\n"’ > rg.txt  samtools merge -rh rg.txt merged.bam ga.bam 454.bam  The value in a **RG** tag is determined by the file name the read is coming from. In this example, in the *merged.bam*, reads from *ga.bam* will be attached *RG:Z:ga*, while reads from *454.bam* will be attached *RG:Z:454*. |
| o | Call SNPs and short INDELs for one diploid individual:  samtools mpileup -ugf ref.fa aln.bam | bcftools view -bvcg - > var.raw.bcf  bcftools view var.raw.bcf | vcfutils.pl varFilter -D 100 > var.flt.vcf  The **-D** option of varFilter controls the maximum read depth, which should be adjusted to about twice the average read depth. One may consider to add **-C50** to **mpileup** if mapping quality is overestimated for reads containing excessive mismatches. Applying this option usually helps **BWA-short** but may not other mappers. |
| o | Generate the consensus sequence for one diploid individual:  samtools mpileup -uf ref.fa aln.bam | bcftools view -cg - | vcfutils.pl vcf2fq > cns.fq |
| o | Call somatic mutations from a pair of samples:  samtools mpileup -DSuf ref.fa aln.bam | bcftools view -bvcgT pair - > var.bcf  In the output INFO field, *CLR* gives the Phred-log ratio between the likelihood by treating the two samples independently, and the likelihood by requiring the genotype to be identical. This *CLR* is effectively a score measuring the confidence of somatic calls. The higher the better. |
| o | Call de novo and somatic mutations from a family trio:  samtools mpileup -DSuf ref.fa aln.bam | bcftools view -bvcgT pair -s samples.txt - > var.bcf  File *samples.txt* should consist of three lines specifying the member and order of samples (in the order of child-father-mother). Similarly, *CLR* gives the Phred-log likelihood ratio with and without the trio constraint. *UGT* shows the most likely genotype configuration without the trio constraint, and *CGT* gives the most likely genotype configuration satisfying the trio constraint. |
| o | Phase one individual:  samtools calmd -AEur aln.bam ref.fa | samtools phase -b prefix - > phase.out  The **calmd** command is used to reduce false heterozygotes around INDELs. |
| o | Call SNPs and short indels for multiple diploid individuals:  samtools mpileup -P ILLUMINA -ugf ref.fa \*.bam | bcftools view -bcvg - > var.raw.bcf  bcftools view var.raw.bcf | vcfutils.pl varFilter -D 2000 > var.flt.vcf  Individuals are identified from the **SM** tags in the **@RG** header lines. Individuals can be pooled in one alignment file; one individual can also be separated into multiple files. The **-P** option specifies that indel candidates should be collected only from read groups with the **@RG-PL** tag set to *ILLUMINA*. Collecting indel candidates from reads sequenced by an indel-prone technology may affect the performance of indel calling. |
| o | Derive the allele frequency spectrum (AFS) on a list of sites from multiple individuals:  samtools mpileup -Igf ref.fa \*.bam > all.bcf  bcftools view -bl sites.list all.bcf > sites.bcf  bcftools view -cGP cond2 sites.bcf > /dev/null 2> sites.1.afs  bcftools view -cGP sites.1.afs sites.bcf > /dev/null 2> sites.2.afs  bcftools view -cGP sites.2.afs sites.bcf > /dev/null 2> sites.3.afs  ......  where *sites.list* contains the list of sites with each line consisting of the reference sequence name and position. The following **bcftools** commands estimate AFS by EM. |
| o | Dump BAQ applied alignment for other SNP callers:  samtools calmd -bAr aln.bam > aln.baq.bam  It adds and corrects the **NM** and **MD** tags at the same time. The **calmd** command also comes with the **-C** option, the same as the one in **pileup** and **mpileup**. Apply if it helps. |
|  |  |

**LIMITATIONS**

|  |  |
| --- | --- |
| o | Unaligned words used in bam\_import.c, bam\_endian.h, bam.c and bam\_aux.c. |
| o | Samtools paired-end rmdup does not work for unpaired reads (e.g. orphan reads or ends mapped to different chromosomes). If this is a concern, please use Picard’s MarkDuplicate which correctly handles these cases, although a little slower. |
|  |  |

**AUTHOR**

Heng Li from the Sanger Institute wrote the C version of samtools. Bob Handsaker from the Broad Institute implemented the BGZF library and Jue Ruan from Beijing Genomics Institute wrote the RAZF library. John Marshall and Petr Danecek contribute to the source code and various people from the 1000 Genomes Project have contributed to the SAM format specification.