FASTQC

USAGE

fastqc [options] seqfile1 seqfile2 .. seqfileN

The (main) options for the program as as follows:

-h | --help	Print help message and exit
-o | --outdir <DIR>	Create all output files in the specified output directory. Please note that this directory must exist as the program will not create it. If this option is not set then the output file for each sequence file is created in the same directory as the sequence file which was processed.
-t | --threads <INT>	Specifies the number of files which can be processed simultaneously. Each thread will be allocated 250MB of memory so you shouldn't run more threads than your available memory will cope with, and not more than 6 threads on a 32 bit machine.
-c | --contaminants <FILE>	Specifies a non-default file which contains the list of contaminants to screen overrepresented sequences against. The file must contain sets of named contaminants in the form name[tab]sequence. Lines prefixed with a hash will be ignored.
-a | --adapters <FILE>	Specifies a non-default file which contains the list of adapter sequences which will be explicity searched against the library. The file must contain sets of named adapters in the form name[tab]sequence. Lines prefixed with a hash will be ignored.
-k | --kmers <INT>	Specifies the length of Kmer to look for in the Kmer content module. Specified Kmer length must be between 2 and 10. Default length is 7 if not specified.

TRIM_GALORE

USAGE:

trim_galore [options] <filename(s)>

-h | --help	Print help message and exit
-q | --quality <INT>	Trim low-quality ends from reads in addition to adapter removal. For RRBS samples, quality trimming will be performed first, and adapter trimming is carried in a second round. Other files are quality and adapter trimmed in a single pass. The algorithm is the same as the one used by BWA (Subtract INT from all qualities; compute partial sums from all indices to the end of the sequence; cut sequence at the index at which the sum is minimal). Default Phred score: 20.
--phred33	Instructs Cutadapt to use ASCII+33 quality scores as Phred scores (Sanger/Illumina 1.9+ encoding) for quality trimming. Default: ON.
--phred64	Instructs Cutadapt to use ASCII+64 quality scores as Phred

	scores (Illumina 1.5 encoding) for quality trimming.
--fastqc	Run FastQC in the default mode on the FastQ file once trimming is complete.
--illumina	Adapter sequence to be trimmed is the first 13bp of the Illumina universal adapter 'AGATCGGAAGAGC' instead of the default auto-detection of adapter sequence.
--gzip	Compress the output file with GZIP. If the input files are GZIP-compressed the output files will automatically be GZIP compressed as well. As of v0.2.8 the compression will take place on the fly.
--length <INT>	Discard reads that became shorter than length INT because of either quality or adapter trimming. A value of '0' effectively disables this behaviour. Default: 20 bp. For paired-end files, both reads of a read-pair need to be longer than <INT> bp to be printed out to validated paired-end files (see option --paired). If only one read became too short there is the possibility of keeping such unpaired single-end reads (see --retain_unpaired). Default pair-cutoff: 20 bp.
--max_n COUNT	The total number of Ns (as integer) a read may contain before it will be removed altogether. In a paired-end setting, either read exceeding this limit will result in the entire pair being removed from the trimmed output files.
--trim-n	Removes Ns from either side of the read.
--paired	This option performs length trimming for paired-end files. To pass the validation test, both sequences of a sequence pair are required to have a certain minimum length which is governed by the option --length (see above). If only one read passes this length threshold the other read can be rescued (see option --retain_unpaired). Using this option lets you discard too short read pairs without disturbing the sequence-by-sequence order of FastQ files which is required by many aligners. Trim Galore! expects paired-end files to be supplied in a pairwise fashion, e.g. file1_1.fq file1_2.fq

BWA INDEX

Usage: bwa index [options] <in.fasta>

Options:

-a STR	BWT construction algorithm: bwtsw or is [auto]
-p STR	prefix of the index [same as fasta name]

BWA MEM

Usage: bwa mem [options] <idxbase> <in1.fq> [in2.fq]

Algorithm options:

-t INT	number of threads [1]
-B INT	penalty for a mismatch [4]
-O INT[,INT]	gap open penalties for deletions and insertions [6,6]
-E INT[,INT]	gap extension penalty; a gap of size k cost '{-O} + {-E}*k' [1,1]
-L INT[,INT]	penalty for 5'- and 3'-end clipping [5,5]
-U INT	penalty for an unpaired read pair [17]

Input/output options:

-T INT	minimum score to output [30]
-a	output all alignments for SE or unpaired PE
-M	mark shorter split hits as secondary

SAMTOOLS

Usage:

samtools <command> [options]

Main commands:

index	Index the alignment
rmdup	Remove PCR duplicates
sort	Sort alignment file
depth	Compute the positional depth
flagstat	Print simple stats
tview	Text alignment viewer
view	Text viewing, filtering and SAM<->BAM<->CRAM conversion

Individual command usage and options

INDEX

samtools index in.bam

NOTE: in.bam needs to be sorted before indexing

RMDUP

samtools rmdup in.bam out.bam

NOTE: in.bam needs to be sorted before removing duplicates

SORT

samtools sort in.bam out.sorted.bam

DEPTH

samtools depth in1.bam [in2.bam […]]

NOTE: All input BAMs need to be sorted before computing depth

FLAGSTAT

samtools flagstat in.bam

TVIEW

samtools tview [-p chr:pos] [-d display] *in.sorted.bam* [*ref.fasta*]

Options:

-d	Output as (H)tml or (C)urses or (T)ext
-p chr:pos	Go directly to this position

VIEW

samtools view [options] in.bam|in.sam

Options:

-b	Output in BAM format
-h	Include header in SAM output
-c	Print only the count of matching records
-o	Output file name [default is printing to stdout]
-f	Only output alignments with all bits set in INT present in the FLAG field
-F	Do not output alignments with all bits set in INT present in the FLAG field i.e. Output alignments with none of the bits set in INT

Examples of using samtools view:

Conversion from SAM (text) to BAM (binary) formats

samtools view -S -h -b –o *output.bam input.sam*

Conversion from BAM (binary) to SAM (binary) formats

samtools view –o *output.sam input.bam*

Visualization of a BAM (binary) file

samtools view -h *input.bam*

Filtering out unmapped reads in BAM files

samtools view -h -F 4 *input.bam* > *filtered.bam*

Extracting unmapped reads from BAM files

samtools view -h -f 4 *input.bam* > *unmapped.bam*

Variant Effect Predictor (VEP)

Run vep locally

vep -i input.txt -o output.txt --offline --dir /vep_dir_path

Full option list:

https://www.ensembl.org/info/docs/tools/vep/script/vep_options.html

Selected options:

-i	Input file name. If not specified, the script will attempt to read from STDIN.
-o	Output file name. The script can write to STDOUT by specifying STDOUT as the output file name - this will force quiet mode. *Default = "variant_effect_output.txt"*
--force	By default, the script will fail with an error if the output file already exists. You can force the overwrite of the existing file by using this flag. *Not used by default*
--offline	Enable offline mode. No database connections will be made, and a cache file or GFF/GTF file is required for annotation. Add --refseq to use the refseq cache (if installed). *Not used by default*
--dir	Specify the base cache/plugin directory to use. *Default = "$HOME/.vep/"*
--everything	Shortcut flag to switch on all of the following: --sift b, --polyphen b, --ccds, --uniprot, --hgvs, --symbol, --numbers, --domains, --regulatory, --canonical, --protein, --biotype, --uniprot, --tsl, --appris, --gene_phenotype --af, --af_1kg, --af_esp, --af_gnomad, --max_af, --pubmed, --variant_class
--symbol	Adds the gene symbol (e.g. HGNC) (where available) to the output. *Not used by default*
--canonical	Adds a flag indicating if the transcript is the canonical transcript for the gene. *Not used by default*
--max_af	Report the highest allele frequency observed in any population from 1000 genomes, ESP or gnomAD. *Not used by default*
--filter__common	Shortcut flag for the filters below - this will exclude variants that have a co-located existing variant with global AF > 0.01 (1%). May be modified using any of the following freq_* filters. *Not used by default*
--vcf	Writes output in VCF format. Consequences are added in the INFO field of the VCF file, using the key "CSQ". Data fields are encoded separated by "|"; the order of fields is written in the VCF header. *Not used by default*

Vcftools

vcftools is a suite of functions for use on genetic variation data in the form of VCF and BCF files.

Reference guide:

http://vcftools.sourceforge.net/man_latest.html

Running example

1) Calculate allele frequence across multiple samples

vcftools --vcf input_file.vcf --freq --out output_file

2) Remove indels

vcftools --vcf input_file.vcf --remove-indels --recode --out output_snps

3) Find differences between two vcfs

vcftools --vcf input_file1.vcf --diff input_file2.vcf --diff-site --out vars1_vars2

Selected options:

--vcf	This option defines the VCF file to be processed.
--out	This option defines the output filename prefix for all files generated by vcftools.
--freq	Outputs the allele frequency for each site in a file with the suffix ".frq".
--remove-indels	Exclude sites that contain an indel.
--recode	These options are used to generate a new file in either VCF or BCF from the input VCF or BCF file after applying the filtering options specified by the user. The output file has the suffix ".recode.vcf"
--diff	This option compare the original input file to this specified VCF.
--diff-site	Outputs the sites that are common / unique to each file. The output file has the suffix ".diff.sites_in_files".
--indv	Specify an individual to be kept from the analysis. This option can be used multiple times to specify multiple individuals.
--chr	Includes or excludes sites with identifiers matching <chromosome>.
--from-bp --to-bp	These options specify a lower bound and upper bound for a range of sites to be processed. Sites with positions less than or greater than these values will be excluded. These options can only be used in conjunction with a single usage of --chr. Using one of these does not require use of the other.

VarScan2

VarScan is a platform-independent mutation caller for targeted, exome, and whole-genome resequencing data.

Reference guide:

http://varscan.sourceforge.net/

Calling pipelines:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4278659/

1) Germline Variant Calling in Individual or Multiple samples

java -jar VarScan.jar [m]pileup2snp sample.[m]pileup [options]

Selected options:

[m]pileup2snp	Identify SNPs from an [m]pileup file.
[m]pileup2indel	Identify indels an [m]pileup file
--min-coverage	Minimum read depth at a position to make a call
--min-var-freq	Minimum variant allele frequency threshold
--min-freq-for-homo	Minimum frequency to call homozygote
--p-value	Default p-value threshold for calling variants.
--output-vcf	If 1 returns the output in vcf format.

2) Somatic Mutation Detection in Tumor-Normal Pairs

java -jar VarScan.jar somatic normal-tumor.mpileup output.basename --mpileup 1

Selected options:

--min-coverage	Minimum read depth at a position to make a call
--min-var-freq	Minimum variant allele frequency threshold
--somatic-p-value	P-value threshold to call a somatic site
--p-value	P-value threshold to call a heterozygote.
--normal-purity	Estimated purity (non-tumor content) of normal sample
--tumor-purity	Estimated purity (tumor content) of tumor sample

```
java -jar VarScan.jar processSomatic output.basename.snp
```

Divide the variants in 3 groups: Somatic, Germline and LOH. For these groups, the subsets of high-confidence variants are determined using a few empirically-derived criteria. For example, high-confidence somatic mutations have tumor VAF>15%, normal VAF<5%, and a somatic p-value of <0.03. The same command can be executed for the indels.

```
java -jar VarScan.jar somaticFilter output.basename.snp.Somatic.hc
    --indel-file output.basename.indel --output-file output.basename.snp.Somatic.hc.filter
```

The above command identifies and removes somatic mutations that are likely false positives due to alignment problems near indels.