Babraham Bioinformatics

May 18, 2016

SNPsplit

v0.3.0

SNPsplit is an allele-specific alignment sorter which is designed to read alignment files in SAM/ BAM format and determine the allelic origin of reads that cover known SNP positions. For this to work a library must have been aligned to a genome which had all SNP positions masked by the ambiguity base ’N’, and aligned using aligners that are capable of using a reference genome which contains ambiguous nucleobases. Examples of supported alignment programs are Bowtie2, Hisat2, Bismark, HiCUP, TopHat or STAR (for some tips using Hisat2 or STAR alignments please see below). In addition, a list of all known SNP positions between the two different genomes must be provided using the option ––snp_file. SNP information to generate N-masked genomes needs to be acquired elsewhere, e.g. for different strains of mice you can find variant call files at the Mouse Genomes Project page at http://www.sanger.ac.uk/resources/mouse/genomes/.

SNPsplit now offers a separate genome preparation step that allows you to generate N-masked (or fully incorporated) SNP genomes for single or dual hybrid strains for all strains of the Mouse Genomes Project. Please see below for further details.

SNPsplit operates in two stages:

I) **SNPsplit-tag:** SNPsplit analyses reads (single-end mode) or read pairs (paired-end mode) for overlaps with known SNP positions, and writes out a tagged BAM file in the same order as the original file. Unsorted paired-end files are sorted by name first.

II) **SNPsplit-sort:** the tagged BAM file is read in again and sorted into allele-specific files. This process may also be run as a stand-alone module on tagged BAM files (tag2sort).

The SNPsplit-tag module determines whether a read can be assigned to a certain allele and appends an additional optional field 'XX:Z:' to each read. The tag can be one of the following:

XX:Z:UA - Unassigned
XX:Z:G1 - Genome 1-specific
XX:Z:G2 - Genome 2-specific

XX:Z:CF - Conflicting

The SNPsplit-sort module **tag2sort** reads in the tagged BAM file and sorts the reads (or read pairs) according to their XX:Z: tag (or the combination of tags for paired-end or Hi-C reads) into sub-files.

SNPsplit workflow in more detail

1) sam2bam Optional. If the supplied file is a SAM file it will first be converted to BAM format (using `samtools view`).

2) Sorting Paired-end files might require the input file to be sorted by read ID before continuing with the allele-tagging (Read 1 and Read 2 of a pair are expected to follow each other in the input BAM file). Unless specifically stated, paired-end BAM files will be sorted by position (using `samtools sort -n`; output file ending in `.sortedByName.bam`). For files that already contain R1 and R2 on two consecutive lanes, the sorting step may be skipped using the option `--no_sort`. Single-end files or Hi-C files generated by `HiCUP` do not require sorting.

3) SNP positions are read in from the SNP file (which may be `GZIP` compressed (ending in `.gz`) or plain text files). The SNP file is expected to be in the following format (tab-delimited):

ID	Chr	Position	SNP value	Ref/SNP
18819008	5	48794752	1	C/T
40491905	11	63643453	1	A/G
44326884	12	96627819	1	T/A

Only the information contained in fields 'Chr (Chromosome)', 'Position' and 'Ref/SNP’ base are being used for analysis. The genome containing the 'Ref' base is used for ‘genome 1 specific reads (G1)’, the genome containing the ‘SNP’ base for ‘genome 2 specific reads (G2)’. If reads do not overlap any SNP positions they are considered ‘Unassigned (UA)’, i.e. they are not informative for one allele or another. In the rare case that a read contains both genome 1- and genome 2-specific base(s), or that the SNP position was deleted the read is regarded as ‘Conflicting (CF)’.

It is probably noteworthy that the determination of overlaps correctly handles the CIGAR operations **M** (match), **D** (deletion in the read), **I** (insertion in the read) and **N** (skipped regions, used for splice mapping by TopHat). Other CIGAR operations are currently not supported.

4) Upon completion, a small allele-specific tagging report is printed to screen and to a report file (.SNPsplit_report.txt) for archiving purposes.

5) Once the tagging has completed, the **tag2sort** module reads in the tagged BAM file and sorts it into various sub-files according to their XX:Z: tag. Both single and paired-end files are sorted into the four categories:

tag UA - Unassigned
tag G1 - Genome 1-specific
tag G2 - Genome 2-specific
tag CF - Conflicting (not reported by default)

6) Upon completion, an allele-specific sorting report is printed out on screen and to a report file for archiving purposes (*.SNPsplit_sort.txt). If the sorting was launched by SNPsplit and not run stand-alone (as **tag2sort**) the sorting report will also be written into the main SNPsplit report (*.SNPsplit_report.txt).

Specific considerations

Paired-end:

In paired-end mode, both reads are used for the classification. Read pairs with conflicting reads (tag CF) or pairs containing both tags G1 and G2 are considered conflicting and are not reported by default. Reporting of these reads can be enabled using the option `--conflicting`.

Singleton alignments in the allele-tagged paired-end file (which is the default for e.g. `TopHat`) are also sorted into the above four files. Specifying `--singletons` will write these alignments to special singleton files instead (ending in *_st.bam).

Hi-C data:

Assumes data processed with `HiCUP` (www.bioinformatics.babraham.ac.uk/projects/hicup/) as input, i.e. the input BAM files are by definition paired-end and Reads 1 and 2 follow each other. Hi-C sorting discriminates several more possible read combinations:

G1-G1
G2-G2
G1-UA
G2-UA
G1-G2
UA-UA

Again, read pairs containing a conflicting read (tag CF) are not printed out by default, but this may be enabled using the option `--conflicting`. For an example report please see below.

RNA-Seq alignments with STAR:

Alignment files produced by the `Spliced Transcripts Alignment to a Reference (STAR)` aligner (https://github.com/alexdobin/STAR/) also work well with SNPsplit, however a few steps need to be adhered to to make this work.

1) Since **SNPsplit** only recognises the CIGAR operations M, I, D and N (see above) alignments need to be run in end-to-end mode and not using local alignments (which may result in soft-clipping). This can be accomplished using the option:
   `'--alignEndsType EndToEnd'`

2) **SNPsplit** requires the MD:Z: field of the BAM alignment to work out mismatches involving masked N positions. Since `STAR` doesn't report the MD:Z: field by default it needs to be instructed to do so, e.g.:

   `'--outSAMattributes NH HI NM MD'`

3) To save some time and avoid having to sort the reads by name, `STAR` can be told to leave R1 and R2 following each other in the BAM file using the option:

   `'--outSAMtype BAM Unsorted'`

Alignments with Hisat2:

DNA or RNA alignment files produced by Hisat2 (https://ccb.jhu.edu/software/hisat2/index.shtml, https://github.com/infphilo/hisat2) also work well with SNPsplit if you make sure that Hisat2 doesn't perform soft-clipping. At the time of writing the current version of Hisat2 (2.0.3-beta) does perform soft-clipping (CIGAR operation: S) even though this is not well documented (or in fact the documentation on Github suggests that the default mode is end-to-end which should not perform any soft-clipping whatsoever). Until the end-to-end mode works as expected users will have to set the penalty for soft-clipping so high that it is effectively not performed (`--sp` is the option governing the soft-clipping penalty). We suggest adding the following option to the Hisat2 command:

`'--sp 1000,1000'`

Bisulfite-Seq data:

This mode assumes input data has been processed with the bisulfite mapping tool `Bismark` (www.bioinformatics.babraham.ac.uk/projects/bismark/). SNPsplit will run a quick check at the start of a run to see if the file provided appears to be a Bismark file, and set the flags `--bisulfite` and/or `--paired` automatically. In addition it will perform a quick check to see if a paired-end file appears to have been positionally sorted, and if not will set the `--no_sort` flag (this data is extracted from the @PG header line). Paired-end (`--paired`) or bisulfite (`--bisulfite`) mode can also be set manually. Paired-end mode requires Read 1 and Read 2 of a pair to follow each other in consecutive lines.

Utilisation of SNP positions and allele assignment of bisulfite reads

In contrast to the standard mode of using all known SNP positions, SNPs involving C to T transitions may not be used for allele-specific sorting since they might reflect either a SNP or a methylation state. This includes all of the following Reference/SNP combinations:

C/T or T/C for forward strand alignments and G/A or A/G for reverse strand alignments.

The number of SNP positions that have been skipped because of this bisulfite ambiguity is reported in the report file. These positions may however be used to assign opposing strand alignments since they do not involve C to T transitions directly. For that reason, the bisulfite call processing also extracts the bisulfite strand information from the alignments in addition to the basecall at the position involved. For any SNPs involving C positions that are not C to T SNPs both methylation states, i.e. C and T, are allowed to match the C position.

For SNPs which are masked by Ns in the genome no methylation call will be performed, i.e. they will receive a '.' (dot) in the methylation call string. This means that SNP positions are used for allele-sorting only but do not participate in calling methylation. While this may reduce the number of total methylation calls somewhat it effectively eliminates the problem of using positions with potentially incorrect methylation status.

SNPsplit genome preparation

SNPsplit_genome_preparation is designed to read in a variant call files from the Mouse Genomes Project (e.g. this latest file: ftp://ftp-mouse.sanger.ac.uk/current_snps/mgp.v5.merged.snps_all.dbSNP142.vcf.gz) and generate new genome versions where the strain SNPs are either incorporated into the new genome (full sequence) or masked by the ambiguity nucleobase 'N' (N-masking).

SNPsplit_genome_preparation may be run in two different modes:

Single strain mode:

1) The VCF file is read and filtered for high-confidence SNPs in the strain specified with `--strain <name>`

2) The reference genome (given with `--reference_genome <genome>`) is read into memory, and the filtered high-confidence SNP positions are incorporated either as N-masking (default) or full sequence (option `--full_sequence`)

Dual strain mode:

1) The VCF file is read and filtered for high-confidence SNPs in the strain specified with `--strain <name>`

2) The reference genome (given with `--reference_genome <genome>`) is read into memory, and the filtered high-confidence SNP positions are incorporated as full sequence and optionally as N-masking

3) The VCF file is read one more time and filtered for high-confidence SNPs in strain 2 specified with `--strain2 <name>`

4) The filtered high-confidence SNP positions of strain 2 are incorporated as full sequence and optionally as N-masking

5) The SNP information of strain and strain 2 relative to the reference genome build are compared, and a new Ref/SNP annotation is constructed whereby the new Ref/SNP information will be Strain/Strain2 (and no longer the standard reference genome strain Black6 (C57BL/6J))

6) The full genome sequence given with `--strain <name>` is read into memory, and the high-confidence SNP positions between Strain and Strain2 are incorporated as full sequence and optionally as N-masking

The resulting `.fa` files are ready to be indexed with your favourite aligner. Proved and tested aligners include Bowtie2, Tophat, STAR, Hisat2, HiCUP and Bismark. Please note that STAR and Hisat2 may require you to disable soft-clipping, please see above for more details

Both the SNP filtering and the genome preparation write out little report files for record keeping.

Examples

Paired-end report (2x50bp):

```
Input file:                                  'FVBNJ_Cast.bam'
Writing allele-flagged output file to:       'FVBNJ_Cast.allele_flagged.bam'


Allele-tagging report
=====================
Processed 194564995 read alignments in total
149380724 reads were unassignable (76.78%)
35143075 reads were specific for genome 1 (18.06%)
9860248 reads were specific for genome 2 (5.07%)
118662 reads did not contain one of the expected bases at known SNP positions (0.06%)
180948 contained conflicting allele-specific SNPs (0.09%)


SNP coverage report
===================
N-containing reads:   45276050
non-N:                149262062
total:                194564995
Reads had a deletion of the N-masked position (and were thus dropped):     26883 (0.01%)
Of which had multiple deletions of N-masked positions within the same read:       30

Of valid N containing reads,
N was present in the list of known SNPs:            61087551 (99.99%)
N was not present in the list of SNPs:              4773 (0.01%)

Input file:                                         'FVBNJ_Cast.allele_flagged.bam'
Writing unassigned reads to:                        'FVBNJ_Cast.unassigned.bam'
Writing genome 1-specific reads to:                 'FVBNJ_Cast.genome1.bam'
Writing genome 2-specific reads to:                 'FVBNJ_Cast.genome2.bam'


Allele-specific paired-end sorting report
=========================================
Read pairs/singletons processed in total:           98215744
        thereof were read pairs:                    96349251
        thereof were singletons:                    1866493
Reads were unassignable (not overlapping SNPs):     61174812 (62.29%)
        thereof were read pairs:     59662537
        thereof were singletons:     1512275
Reads were specific for genome 1:                   28657857 (29.18%)
        thereof were read pairs:     28446094
        thereof were singletons:     211763
Reads were specific for genome 2:                   8122687 (8.27%)
        thereof were read pairs:     7985424
        thereof were singletons:     137263
Reads contained conflicting SNP information:        260388 (0.27%)
        thereof were read pairs:     255196
        thereof were singletons:     5192
```

Hi-C report (2x100bp):

```
Input file:                                  Black6_129S1.bam
Writing allele-flagged output file to:       Black6_129S1.allele_flagged.bam


Allele-tagging report
=====================
Processed 94887256 read alignments in total
59662038 reads were unassignable (62.88%)
19851697 reads were specific for genome 1 (20.92%)
15047281 reads were specific for genome 2 (15.86%)
47261 reads did not contain one of the expected bases at known SNP positions (0.05%)
326240 contained conflicting allele-specific SNPs (0.34%)


SNP coverage report
===================
N-containing reads:    35231977
non-N:                 59614777
total:                 94887256
Reads had a deletion of the N-masked position (and were thus dropped):     40502 (0.04%)
Of which had multiple deletions of N-masked positions within the same read:       59

Of valid N containing reads,
N was present in the list of known SNPs:     57101748 (99.99%)
N was not present in the list of SNPs:       4211 (0.01%)

Input file:                                          Black6_129S1.allele_flagged.bam'
Writing unassigned reads to:                         Black6_129S1.UA_UA.bam'
Writing genome 1-specific reads to:                  Black6_129S1.G1_G1.bam'
Writing genome 2-specific reads to:                  Black6_129S1.G2_G2.bam'
Writing G1/UA reads to:                              Black6_129S1.G1_UA.bam'
Writing G2/UA reads to:                              Black6_129S1.G2_UA.bam'
Writing G1/G2 reads to:                              Black6_129S1.G1_G2.bam'


Allele-specific paired-end sorting report
=========================================
Read pairs processed in total:                       47443628
Read pairs were unassignable (UA/UA):                18862725 (39.76%)
Read pairs were specific for genome 1 (G1/G1):       3533932 (7.45%)
Read pairs were specific for genome 2 (G2/G2):       2592040 (5.46%)
Read pairs were a mix of G1 and UA:                  12306421 (25.94%). Of these,
                       were G1/UA: 6018598
                       were UA/G1: 6287823
Read pairs were a mix of G2 and UA:                  9430675 (19.88%). Of these,
                       were G2/UA: 4603429
                       were UA/G2: 4827246
Read pairs were a mix of G1 and G2:                  395296 (0.83%). Of these,
                       were G1/G2: 198330
                       were G2/G1: 196966
Read pairs contained conflicting SNP information:    322539 (0.68%)
```

BS-Seq report (2x100bp):

```
Input file:                                  '129_Cast_bismark_bt2_pe.bam'
Writing allele-flagged output file to:       '129_Cast_bismark_bt2_pe.allele_flagged.bam'


Allele-tagging report
=====================
Processed 162441396 read alignments in total
Reads were unaligned and hence skipped: 0 (0.00%)
109109113 reads were unassignable (67.17%)
30267901 reads were specific for genome 1 (18.63%)
22697499 reads were specific for genome 2 (13.97%)
15807753 reads did not contain one of the expected bases at known SNP positions (9.73%)
366883 contained conflicting allele-specific SNPs (0.23%)


SNP coverage report
===================
SNP annotation file:   ../all_Cast_SNPs_129S1_reference.mgp.v4.txt.gz
N-containing reads:    68984287
non-N:                 93301360
total:                 162441396
Reads had a deletion of the N-masked position (and were thus dropped):     155749 (0.10%)
Of which had multiple deletions of N-masked positions within the same read:       65

Of valid N containing reads,
N was present in the list of known SNPs:     119119643 (99.99%)
Positions were skipped since they involved C>T SNPs: 38464451
N was not present in the list of SNPs:               7517 (0.01%)

Input file:
        129_Cast_bismark_bt2_pe.allele_flagged.bam'
Writing unassigned reads to:                         129_Cast_bismark_bt2_pe.unassigned.bam'
Writing genome 1-specific reads to:                  129_Cast_bismark_bt2_pe.genome1.bam'
Writing genome 2-specific reads to:                  129_Cast_bismark_bt2_pe.genome2.bam'


Allele-specific paired-end sorting report
=========================================
Read pairs/singletons processed in total:            81220698
        thereof were read pairs:                     81220698
        thereof were singletons:                     0
Reads were unassignable (not overlapping SNPs):             40420625 (49.77%)
        thereof were read pairs:      40420625
        thereof were singletons:      0
Reads were specific for genome 1:                    23037433 (28.36%)
        thereof were read pairs:      23037433
        thereof were singletons:      0
Reads were specific for genome 2:                    17303663 (21.30%)
        thereof were read pairs:      17303663
        thereof were singletons:      0
Reads contained conflicting SNP information:         458977 (0.57%)
        thereof were read pairs:      458977
        thereof were singletons:      0
```

Full list of options for SNPsplit

USAGE: `SNPsplit [options] --snp_file <SNP.file.gz> [input file(s)]`

`Input file(s)`

Mapping output file in SAM or BAM format. SAM files (ending in .sam) will first be converted to BAM files.

`--snp_file`

Mandatory file specifying SNP positions to be considered, may be a plain text file of `gzip` compressed. Currently, the SNP file is expected to be in the following format:

SNP-ID	Chromosome	Position	Strand	Ref/SNP
33941939	9	68878541	1	T/G

Only the information contained in fields 'Chromosome', 'Position' and 'Ref/SNP base' are being used for analysis. The genome referred to as 'Ref' will be used as genome 1, the genome containing the 'SNP' base as genome 2.

`--paired`

Paired-end mode. (Default: OFF).

`--singletons`

If the allele-tagged paired-end file also contains singleton alignments (which is the default for e.g. TopHat), these will be written out to extra files (ending in _st.bam) instead of writing everything to combined paired-end and singleton files. Default: OFF.

`--no_sort`

This option skips the sorting step if BAM files are already sorted by read name (e.g. Hi-C files generated by HiCUP). Please note that setting --no_sort for unsorted paired-end files will break the tagging process!

`--hic`

Assumes Hi-C data processed with `HiCUP` (www.bioinformatics.babraham.ac.uk/projects/hicup/) as input, i.e. the input BAM file is paired-end and Reads 1 and 2 follow each other. Thus, this option also sets the flags `--paired` and `--no_sort`. Default: OFF.

`--bisulfite`

Assumes Bisulfite-Seq data processed with Bismark (www.bioinformatics.babraham.ac.uk/projects/bismark/) as input. In paired-end mode (`--paired`), Read 1 and Read 2 of a pair are expected to follow each other in consecutive lines. SNPsplit will run a quick check at the start of a run to see if the provided file appears to be a Bismark file, and set the flags `--bisulfite` and/or `–paired` automatically. In addition it will perform a quick check to see if a paired-end file appears to have been positionally sorted, and if not will set the flag `--no_sort`.

`--samtools-path` The path to your Samtools installation, e.g. /home/user/samtools/. Does not need to be specified explicitly if Samtools is in the PATH already.

SNPsplit-sort specific options (**tag2sort**):

`--sam` The output will be written out in SAM format instead of the default BAM format. SNPsplit will attempt to use the path to Samtools that was specified with `--samtools_path`, or, if it hasn't been specified, attempt to find Samtools in the `PATH` environment.

`--conflicting/--weird` Reads or read pairs that were classified as 'Conflicting' (XX:Z:CF) will be written to an extra file (ending in .conflicting.bam) instead of being simply skipped. Reads may be classified as 'Conflicting' if a single read contains SNP information for both genomes at the same time, or if the SNP position was deleted from the read. Read-pairs are considered '.Conflicting' if either read is was tagged with the XX:Z:CF flag. Default: OFF.

`--help` Displays this help information and exits.

`--verbose` Verbose output (for debugging).

`--version` Displays version information and exits.

Full list of options for SNPsplit_genome_preparation

USAGE: SNPsplit_genome_preparation [options] --vcf_file <file> --reference_genome /path/to/genome/ --strain <strain_name>

`--vcf_file <file>`

Mandatory file specifying SNP information for mouse strains from the Mouse Genomes Project (http://www.sanger.ac.uk/science/data/mouse-genomes-project). The file approved and tested is called 'mgp.v5.merged.snps_all.dbSNP142.vcf.gz'. Please note that future versions of this file or entirely different VCF files might not work out-of-the-box but may require some tweaking. SNP calls are read from the VCF files, and high confidence SNPs are written into a folder in the current working directory called SNPs_<strain_name>/chr<chromosome>.txt, in the following format:

	SNP-ID	Chromosome	Position	Strand	Ref/SNP
example:	33941939	9	68878541	1	T/G

`--strain <strain_name>`

The strain you would like to use as SNP (ALT) genome. Mandatory. For an overview of strain names just run SNPsplit_genome_preparation selecting '`--list_strains`'.

`--list_strains`

Displays a list of strain names present in the VCF file for use with '`--strain <strain_name>`'.

`--dual_hybrid`

Optional. The resulting genome will no longer relate to the original reference specified with '`--reference_genome`'. Instead the new Reference (Ref) is defined by '`--strain <name>`' and the new SNP genome is defined by '`--strain2 <strain_name>`'. '`--dual_hybrid`' automatically sets '`--full_sequence`'.

This will invoke a multi-step process:
1) Read/filter SNPs for first strain (specified with '`--strain < name>`')
2) Write full SNP incorporated (and optionally N-masked) genome sequence for first strain

3) Read/filter SNPs for second strain (specified with `'--strain2 <name>'`)
4) Write full SNP incorporated (and optionally N-masked) genome sequence for second strain
5) Generate new Ref/Alt SNP annotations for Strain1/Strain2
6) Set first strain as new reference genome and construct full SNP incorporated (and optionally N-masked) genome sequences for Strain1/Strain2

`--strain2 <strain_name>` Optional for constructing dual hybrid genomes (see `'--dual_hybrid'` for more information). For an overview of strain names just run `SNPsplit_genome_preparation` selecting `'--list_strains'`.

`--reference_genome` The path to the reference genome, typically the strain 'Black6' (C57BL/6J), e.g. `'--reference_genome /bi/scratch/Genomes/Mouse/GRCm38/'`. Expects one or more FastA files in this folder (file extension: `.fa` or `.fasta`).

`--skip_filtering` This option skips reading and filtering the VCF file. This assumes that a folder named 'SNPs_<Strain_Name>' exists in the working directory, and that text files with SNP information are contained therein in the following format:

	SNP-ID	Chromosome	Position	Strand	Ref/SNP
example:	33941939	9	68878541	1	T/G

`--nmasking` Write out a genome version for the strain specified where Ref bases are replaced with 'N'. In the Ref/SNP example T/G the N-masked genome would now carry an N instead of the T. The N-masked genome is written to a folder called '<strain_name>_N-masked/'. Default: ON.

`--full_sequence` Write out a genome version for the strain specified where Ref bases are replaced with the SNP base. In the Ref/SNP example T/G the full sequence genome would now carry a G instead of the T. The full sequence genome is written out to folder called '<strain_name>_full_sequence/'. May be set in addition to `'--nmasking'`. Default: OFF.

`--no_nmasking` Disable N-masking if it is not desirable. Will automatically set `'--full_sequence'` instead.

`--genome_build <name>` Name of the mouse genome build, e.g. mm10. Will be incorporated into some of the output files. Defaults to 'GRCm38'.

`--help` Displays this help information and exits.

`--version` Displays version information and exits.