SUPERmerge software manual

SUPERmerge takes as input a BAM file and returns as output:

* `<file>.results` = final results table with annotation to the closest related feature that was given in the gtf file. GTF file was filtered based on "gene" in the second column. The output for this is as follows:
    * Interval_chr
    * Interval_start
    * Interval_stop
    * Percentage of bases meeting cutoff criteria
    * Percentage of interval MINUS the extensions that meet cutoff criteria
    * Total size of the interval region minus cutoff criteria
    * GTF_chr (column 1 of the gtf file)
    * GTF_start (column 4 of the gtf file)
    * GTF_stop (column 5 of the gtf file)
    * GTF_annotation (column 9 of the gtf file)
    * Distance to gtf element
    * Additional column with the total read count for a given interval

* `<file>.pdf` = 4 plots produced in the R programming language (see manuscript for details)

SUPERmerge is composed of three primary parameters that are specified as user flags:

1) -d: depth
2) -i: interval grouping factor
3) -g: GTF annotation

Depth represents read pileup at a given genomic location (e.g., "-d 20" returns all single base pair positions in a BAM file harboring at least 20 or more reads). The interval grouping factor (-i) represents a genomic distance (in bp) between any two consecutive positions of sufficient depth (-d), for which any two neighboring base pair positions are considered to be part of the same genomic interval if they are separated by no more than this distance (-i). In practice, this translates to merging any base pair positions that are of sufficient depth -d into one contiguous interval of coverage, as long as the gap between these genomic positions is no greater than -i. We

refer to these contiguous regions of coverage as "coverage islands", since there exists no base pair position within a -i radius of the island (in either the 3-prime or 5-prime end) that is of the sufficient depth -d. However, within any given coverage island, there will always be a certain percentage of base pairs that do not meet the sufficient depth, simply by virtue of the fact that any two consecutive regions of sufficient depth are merged into a continuous interval whenever there is a gap of less than -i bp between them.

As such, these two thresholds (-d and -i) constitute the cutoff parameters by which a user can explore how reads assemble into coverage islands at various levels of exploratory data analysis. Coverage islands are defined as contiguous genomic regions with a certain percentage of read coverage above a set threshold depth and a distance of at least -i bp from their nearest neighboring island. Once these coverage islands are calculated, the -g flag specifies a gene transfer format (GTF) file that annotates them with their respective genomic features (see manuscript for details).

An example of a full command issued from the command-line is:

./supermerge -d 20 -i 500 -g gencodeV19annotation.gtf file.bam

This command would be repeated by the user across all BAM files in the sample pool, for any desired combinations of -d and -i.