PRESM user's manual

PRESM stands for **P**ersonal **R**eference **E**ditor for **S**omatic **M**utation discovery. Compared with other reference genome editor software which apply diploid reference genome to improve read mapping accuracy, PRESM provides two haploid reference genomes. The pipeline of PRESM involves three steps: Firstly, germline mutations are discovered by another tool, e.g., GATK, and are used to make personalized references to call somatic mutations. Secondly, a reference genome composed of all personal variants (including both heterozygous and homozygous sites) is used as "decoy" to capture the heterozygous variants in reads. Thirdly, PRESM changes the reads by replacing all heterozygous alleles with the corresponding reference alleles and maps the modified reads back to another personalized reference genome that contains only homozygous changes. The output of this step is a BAM file ready for any somatic mutation callers to use. We intend to offer long-term maintenance for PRESM and continue adding our new functions into it.

1. INSTALLATION

The PRESM is a batteries-included executable, therefore no installion is needed. Just copy the executable presm.jar and run it using the standard command for java package:

java –jar presm.jar

2. Functions

1. Processing variant call files generated by GATK, Pindel or other variant call software, i.e., combining two variant call files; homozygous variants or heterozygous variants selecting; variants deduplication.
2. Generating the personalized reference genome according to the germline mutations provided by the users.
3. Generating the personalized vcf files according to the germline mutations provided by the users, for example, the personalized dbsnp.vcf, db.indel.vcf, cosmic.vcf should be generated before the somatic mutation calling procedure of the current methods.
4. Mapping the somatic variants called using personalized reference genome to the universal reference genome.
5. Replacing the alternative alleles with reference bases according to the heterozygous variants provided by the users.

3. Commands and options.

All the functions are used as:

java [-Xmx] –jar /path/to/presm.jar <options>

1. CombineVariants: Combine two variant call files according to the reference genome.

Usage: -F CombineVariants –R ref.fasta –variant1 input1.vcf –variant2 input2.vcf –O output.vcf

–R: input the reference genome file

-variant1: input variant file 1 (in vcf foramt)

-variant2: input variant file 2 (in vcf foramt)

-O: output the combined variant call file in vcf format

2. SelectGenotype: Select homozygous or heterozygous variants in the variant call file provided by the users.

Usage: -F SelectGenotype –genotype homo[heter] –variants input.vcf –O output.vcf

-genotype: Specify the genotype of the variants (homozygous/ heterozygous variants)

-variants: input the variants in vcf format

-O: output the specified genotype variants in vcf format

3. RemoveOverlaps : Remove overlapping variants in variant call file

Usage: -F RemoveOverlaps –R ref.fasta –variants input.vcf –O output.vcf

–R: input the reference genome file

-variants: input the variant in vcf format

-O: output the duplicated variant in vcf format

4. SortVariants: Sort variants according to the reference genome

Usage: -F SortVariants –R ref.fasta –variants input.vcf –O output.vcf

–R: input the reference genome file

-variants: input the variant in vcf format

-O: output the sorted variant in vcf format

5. MakePersonalizedReference: Generate personalized reference genome according to the germline mutations provided by the users.

Usage: -F MakePersonalizedReference –I ref.fasta –germlinemutations input.vcf –O output.fa [–intervals input.intervals] [-genotype home/ heter]

–I: input the reference genome file

-germlinemutations: input the germline mutations in vcf format

-O: output the personalized reference genome in fasta format

Options:

-intervals: specify the region of variants

-genotype: specify the genotype of variants

6. MakePersonalizedVariants: Generate personalized vcf files according to the germline mutations provided by the users.

Usage: -F MakePersonalizedVariants –I input.vcf –O output.vcf –variants variant.vcf [–intervals input.intervals] [-genotype home/ heter] [-removeduplicates]

-I: input the variants in vcf format

-O: output the personalized variants in vcf format

-variants: input the mutations in vcf format

Options:

-intervals: specify the region of variants

-genotype: specify the genotype of variants

-removeduplicates: remove duplicated variants

7. MapVariants: Map the variants from being mapped to the personalized reference genome to being mapped to the universal reference genome.

Usage: -F MapVariants –R ref.fasta –I input.vcf –O output.vcf –germlinemutations variant.vcf [–intervals input.intervals] [-genotype home/ heter] [-removeduplicates]

–R: input the universal reference genome file

-I: input the somatic mutations in vcf format

-O: output the somatic mutations being mapped to the universal reference genome in vcf format

-germlinemutations: input the germline mutations in vcf format

Options:

-intervals: specify the region of variants

-genotype: specify the genotype of variants

-removeduplicates: remove duplicated variants

8. ReplaceGenotype: Replacing alternative alleles with reference bases according to the heterozygous variants provided by the users.

Usage: -F ReplaceGenotype –R ref.fasta –I input.sam –germlinemutations germlinemutations.vcf –O output.sam –readlength len [-genotype home/ heter] [–intervals input.intervals]

–R: input the reference genome file

-I: input the sequence alignment map file in sam format

-variant: input the germline mutations in vcf format

-O: output the replaced sequence alignment map file in sam format

–readlength: the sequencing read length

Options:

-genotype: specify the genotype of variants

-intervals: specify the region of variants

9. ViewFasta: View specified region of sequence in reference genome.

Usage: -F ViewFasta –R ref.fasta [–L input.list] [-region specified region]

–R: input the reference genome file

-L: input the specified region list file, this function was used for viewing multiple regions in the chromosome

-region: input one specified region, this function was used for viewing single region in the chromosome

Example of region specifications format:

chr1   Output whole sequence of chromosome 1 in the reference genome.

chr2:5000   Output the chromosome 2 sequence which begins at base position 5000 and ends at the end of chromosome 2.

chr3:500-600   Output the chromosome 3 sequence which begins at base position 500 and ends at base position 600 of chromosome 3.