Analysis workflow for detection of genome-wide variations in TASUKE+ of RAP-DB

Genome-wide variation data among >500 accessions was provided in the multiple genome browser TASUKE+ of RAP-DB. Here we provide the analysis workflow for detection of the genome-wide variations (SNPs/InDels).

Reference data

- Genome sequences
  - IRGSP-1.0 genome (including organella and unanchored contig sequences)
- Gene annotation
  - RAP-DB (both representative and predicted genes as of 26 Nov 2018)
  - MSU (all genes in RGAP 7)
- Illumina adapter sequences attached to the Trimmomatic program
- Chromosome information (chromosome_list.csv) for TASUKE+

```
$ cat chromosome_list.csv
chr01,43270923,16610866,17243770
chr02,35937250,13541821,13872411
chr03,36413819,19431743,19745569
chr04,35502694,9744480,9973218
chr05,29958434,12390387,12627019
chr06,31248787,15332004,15555636
chr07,29697621,11887856,12272916
chr08,28443022,12847483,13061068
chr09,23012720,2749793,3043847
chr10,23207287,8082722,8309866
chr11,29021106,12039480,12482616
chr12,27531856,11761737,12103486
```

Analysis tools

- Java (JDK 1.8.0_191)
- FastQC v0.11.8
- BWA (bwa v0.7.17)
- SamTools (v1.9)
- BamTools (as of 4 Dec 2018)
- GATK (v4.0.11.0)
- Trimmomatic (v0.38)
- Picard (v2.18.17)
- SnpEff (v4.3t)
- TASUKE+ (version 20190826)

Commands and parameters used in the workflow

1. **Preprocessing of Illumina paired-end reads**

```
$ java -jar trimmomatic-0.38.jar PE \
    -phred33 read.r1.fastq.gz read.r2.fastq.gz \
    read.pe.r1.fastq.gz read.se.r1.fastq.gz read.pe.r2.fastq.gz read.se.r2.fastq.gz \
    ILLUMINACLIP:adapters.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:30
```

2. **Making index of the genome**

```
$ bwa index genome.fa
$ samtools faidx genome.fa
$ java -jar picard.jar CreateSequenceDictionary \
    REFERENCE=genome.fa \
    OUTPUT=genome.dict
```

3. **Alignment of Illumina reads to the reference genome**

```
$ bwa mem -M genome.fa read.pe.r1.fastq.gz read.pe.r2.fastq.gz \
    | samtools sort -o alignment.sort.bam -
$ samtools index alignment.sort.bam
```

4. **Make clean BAM**

```
$ java -jar picard.jar FastqToSam \
    FASTQ=read.pe.r1.fastq.gz \
    FASTQ2=read.pe.r2.fastq.gz \
    OUTPUT=uBAM.bam \
    READ_GROUP_NAME=${SAMPLE_ID} \
    SAMPLE_NAME=${SAMPLE_ID} \
    LIBRARY_NAME=${SAMPLE_ID} \

$ java -jar picard.jar MergeBamAlignment \
    ALIGNED=alignment.sort.bam \
    UNMAPPED=uBAM.bam \
    OUTPUT=alignment.merge.bam \
    REFERENCE_SEQUENCE=genome.fa
```

*User can specify any sample ID in the variable “${SAMPLE_ID}”.

6. **Remove PCR duplicates**

```
$ java -jar picard.jar MarkDuplicates \
    INPUT=alignment.merge.bam \
    OUTPUT=alignment.rmdup.bam \
    METRICS_FILE=rmdup.matrix \
    REMOVE_DUPLICATES=true \
    MAX_RECORDS_IN_RAM=1000000 \
    TMP_DIR=./tmp

$ samtools index alignment.rmdup.bam
```

7. **Variant detection and filtering by GATK**

```
$ gatk HaplotypeCaller \
    --input alignment.rmdup.bam \
    --output variants.g.vcf.gz \
    --reference genome.fa \
    -max-alternate-alleles 2 \
    --emit-ref-confidence GVCF
$ gatk GenotypeGVCFs \
    --variant variants.g.vcf.gz \
    --output variants.genotype.vcf.gz \
    --reference genome.fa

$ gatk VariantFiltration \
    --reference genome.fa \
    --variant variants.genotype.vcf.gz \
    --output variants.filter.genotype.vcf.gz \
    --filter-expression "QD < 5.0 || FS > 50.0 || SOR > 3.0 || MQ < 50.0 || MQRankSum < -2.5 || ReadPosRankSum < -1.0 || ReadPosRankSum > 3.5" \
    --filter-name "FILTER"

$ gatk SelectVariants \
    --reference genome.fa \
    --variant variants.filter.genotype.vcf.gz \
    --output variants.varonly.vcf.gz \
    --exclude-filtered \
    --select-type-to-include SNP \
    --select-type-to-include INDEL
```

8. **Run SnpEff**

```
$ java -jar snpEff.jar build -gtf22 -v RAP_MSU_on_IRGSP-1.0


$ java -jar snpEff.jar \
  -v RAP_MSU_on_IRGSP-1.0 variants.varonly.vcf.gz | gzip -c > variants.snpEff.varonly.vcf.gz
```

*Reference genome sequence (FASTA), gene annotation data (GTF) and protein sequences of all genes (FASTA) should be placed in appropriate directories before running SnpEff.

9. **Make depth data for TASUKE+**

```
$ perl tasuke_bamtodepth.pl \
    -s samtools -c chromosome_list.csv -i alignment.rmdup.bam \
    -o alignment_depth.tsv
```

10. **Upload variation and depth data to TASUKE+**

The following two output files are uploaded to the TASUKE+ database

- variants.snpEff.varonly.vcf.gz
- alignment_depth.tsv