Jun 21, 2017

Structure variation detection

GigaScience

DOI
dx.doi.org/10.17504/protocols.io.gr4bv8w

BGI

Haoxiang Lin

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.gr4bv8w

External link: https://doi.org/10.1093/gigascience/gix067

Protocol Citation: Haoxiang Lin: Structure variation detection. **protocols.io** https://dx.doi.org/10.17504/protocols.io.gr4bv8w

Manuscript citation:
Lan T, Lin H, Zhu W, Laurent TCAM, Yang M, Liu X, Wang J, Wang J, Yang H, Xu X, Guo X, Deep whole-genome sequencing of 90 Han Chinese genomes. GigaScience 6(9). doi: 10.1093/gigascience/gix067

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

Created: December 15, 2016

Last Modified: March 27, 2018

Protocol Integer ID: 4636

Quick align by BWA

1 Quick align by BWA

Software

BWA NAME

Linux OS

Dataset

Contig fasta NAME

Exact align by LASTZ

2 Exact align by LASTZ

Software

LASTZ NAME

LINUX OS

Dataset

Contig fasta NAME

Command

```
lastz --targetcapsule=$CAPSULE $FASTA[nameparse=darkspace] --
strand=both --chain --ambiguous=iupac --gapped --ydrop=50000 --
gap=1000,1 --format=axt --output=$AXT --markend
```

Expected result

LASTZ aligment

Call SV by SOAPSV

3 Call SV by SOAPSV. SOAPSV is a huge pipeline, include many programes and scripts. Several commands lines of key steps are showed.

Software

SOAPsv	NAME
LINUX	OS

Dataset

LASTZ aligment	NAME

Command

```
axtSort $AXT > $SORT_AXT
...
# find best hit in alignments, alignment linearization
best_hit $SORT_AXT > $BEST_AXT
...
intro_indel_1.3 $FINAL_AXT > $SV
```

Call SV by Pindel

4 Call SV by Pindel

Software

PINDEL	NAME
LINUX	OS

Dataset

Merged BAM	NAME

Command

```
pindel -f $HG19 -i $CFG -o $OUT_PREFIX
pindel2vcf -P $PREFIX -r $HG19 -R hg19 -d hg19 -v $VCF
```

Expected result

SV result

Call SV by CNVnator

5 Call SV by CNVnator

Software

cnvnator NAME

LINUX OS

Dataset

Merged BAM NAME

Command

```
./cnvnator -genome hg19 -root out.root -tree $BAM
./cnvnator -genome hg19 -root out.root -his 100
./cnvnator -root out.root -stat 100
./cnvnator -root out.root -partition 100
./cnvnator -root out.root -call 100
```

Expected result

SV result

Call SV by Breakdancer

6 Call SV by Breakdancer

Software

Breakdancer-max NAME

LINUX OS

Dataset

Merged BAM NAME

Command

```
bam2cfg.pl -q 20 -c 3 -g -h $BAM > $CFG
breakdancer -o $PREFIX -q 20 -d $CTX -a -y 30 $CFG
```

Expected result

SV result

Call SV by Genome STRIP

7 Call SV by Genome STRIP

Software

Genome STRIP	NAME
LINUX	OS

Dataset

Merged BAM	NAME

Command

```
java -cp ${classpath} ${mx} \
org.broadinstitute.sting.queue.QCommandLine \ -S
${SV_DIR}/qscript/SVPreprocess.q \ -S ${SV_DIR}/qscript/SVQScript.q \
-gatk ${SV_DIR}/lib/gatk/GenomeAnalysisTK.jar \ -cp ${classpath} \ -
configFile conf/genstrip_parameters.txt \ -disableGATKTraversal \ -
tempDir ${SV_TMPDIR} \ -R $HG19 \ -computeGCProfiles \ -
genomeMaskFile hg19.mask.101.fasta \ -ploidyMapFile hg19.ploidy.map \
-copyNumberMaskFile cn2_mask_hg19.fasta \ -genderMapFile gender.list
\ -runDirectory ${runDir} \ -computeGCProfiles \ -md
${runDir}/metadata \ -jobLogDir ${runDir}/logs \ -I ${bam} \ --
disableJobReport \ -run || exit 1
java -cp ${classpath} ${mx} \
org.broadinstitute.sting.queue.QCommandLine \ -S
${SV_DIR}/qscript/SVDiscovery.q \ -S ${SV_DIR}/qscript/SVQScript.q \ -
gatk ${SV_DIR}/lib/gatk/GenomeAnalysisTK.jar \ --disableJobReport \ -
cp ${classpath} \ -configFile ./genstrip_parameters.txt \ -tempDir
${SV_TMPDIR} \ -R $HG19 \ -genomeMaskFile hg19.mask.101.fasta \ -
genderMapFile gender.list \ -runDirectory ${runDir} \ -md
${runDir}/metadata \ -disableGATKTraversal \ -jobLogDir
${runDir}/logs \ -minimumSize 50 \ -maximumSize 1000000 \ -windowSize
20000000 \ -windowPadding 10000 \ -I ${bam} \ -O ${sites} \ -P
select.validateReadPairs:false \ -run || exit 1
```

Expected result

SV Result

Combine Deletion

8 Combine deletion in individual level between different methods and Combine SV in popluation leverl in different individuals with using in-house scrtips. The methods are similar in 1000 genome paper. Merge exact breakpoint by locations and merge imprecise breakpoint by confident region.

Genotyping by Genome STRIP

9 Genotyping deletions by Genome STRIP

Software

NAME	OS
Genome STRIP	LINUX

Dataset

NAME
Merged BAM and site VCF

Command

```
java -cp ${classpath} ${mx} \
org.broadinstitute.sting.queue.QCommandLine \ -S
${SV_DIR}/qscript/SVGenotyper.q \ -S ${SV_DIR}/qscript/SVQScript.q \ -
gatk ${SV_DIR}/lib/gatk/GenomeAnalysisTK.jar \ --disableJobReport \ -
cp ${classpath} \ -configFile genstrip_parameters.txt \ -tempDir
${SV_TMPDIR} \ -R $HG19 \ -genomeMaskFile hg19.mask.101.fasta \ -
genderMapFile gender.list \ -runDirectory ${runDir} \ -md
${runDir}/metadata \ -jobLogDir ${runDir}/logs \ -I ${bam} \ -vcf
${sites} \ -disableGATKTraversal \ -O ${genotypes} \ -run || exit 1
```

Expected result

SV genotypes in VCF format