Structural Variant Calling for Whole Genome Sequences

Atma Ivancevic atma.ivancevic@adelaide.edu.au
Date created: March 28, 2018
Date updated: March 29, 2018

WARNING: This pipeline is a work in progress. Some things may break as things change. Please let me know if this happens.

Protocol Requirements:

- This protocol is designed for use on a high-performance computer with SLURM queuing (e.g. wiki.adelaide.edu.au/hpc). Scripts can be modified for use on a sufficiently powerful local computer.
- You will need WGS bam files and the human reference genome.
  - We currently use hg19.

Step 1: Set up your files and work environment

- This pipeline assumes all of your BAMs are in the same directory. If this is not the case, consider creating a new directory with symlinks to your BAMs. E.g.:

```
mkdir /data/genomes
cd /data/genomes
ln -s /path/to/bam link_name
```

- All BAM files need to be indexed. This can be done with SAMtools.

```
samtools index bam_file
```

To quickly index many BAMs, try using indexBAM.sh.

- Set up folders for your output and log files. You can do this however you like.

  This is how I set mine up:

```
# General output dir:
$FASTDIR/outputs

# Output dir for SV calling:
$FASTDIR/outputs/SVcalling

# Output dirs for each SV caller:
$FASTDIR/outputs/SVcalling/dellyOut
$FASTDIR/outputs/SVcalling/lumpyOut
$FASTDIR/outputs/SVcalling/mantaOut

# Dir for SLURM output:
$FASTDIR/slurmOUT
# Dir for SLURM logs:
$FASTDIR/slurmLOG
```

- Download and install SV callers. Best practice is to use the latest stable version of each tool. You can put executables in your PATH, or redirect to a dedicated "executables" directory.
  - Delly
  - Lumpy
  - Manta

Step 2: Set off SV calling phase for each caller

- Delly first. An example command for SV calling is:

```
delly call -g hg19.fa -o s1.bcf -x hg19.excl sample.bam
```

Use dellyCalling.sh to run on many genomes. You will need the reference genome (e.g. ucsc.hg19.fasta) and tsv file of regions to exclude (ucsc.hg19.excl.tsv).