PVCTools

（Parallel Variation Calling Tools）

User manual

Contents

1. Introduction

1.1 What is PVCTools

PVCTools is Parallel Variation Calling Tools, which tries to call variation using multiple threads. Basically, the reference genome will be split into small pieces, and corresponding alignment BAM files will be extracted. Under this way, it will speed up the process of variation calling a lot.

1.2 How to cite PVCTools

- Jin J, Liu L, Li Z, Lu P, Xu Y, Cao P[#]&Hu D[#]: PVCTools: Parallel Variation Calling Tools. 2017. Submit to Bioinformatics.

2. Prerequisites

- Samtools: http://www.htslib.org/doc/samtools.html
- GATK: https://software.broadinstitute.org/gatk/
- Picard tools: http://broadinstitute.github.io/picard/
- bamUtil: http://genome.sph.umich.edu/wiki/BamUtil
- bcftools: https://samtools.github.io/bcftools/bcftools.html
- freebayes: https://github.com/ekg/freebayes

3. Source code

- Download: https://github.com/CNaibon/PVCTools/
- Test data: http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/

4. Install

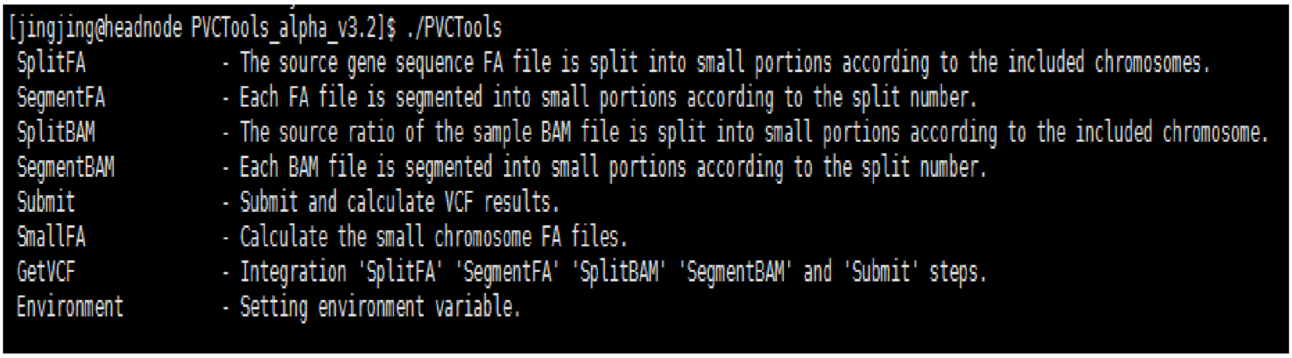
To install PVCTools, please follow:

- unzip PVCTools_alpha_v*.zip
- make

5. Running PVCTools

5.1 Execute

By ./PVCTools, one can see all the parameters in PVCTools. From the following figure, we can see there are 9 modules in PVCTools: SplitFA, SegmentFA, SplitBAM, SegmentBAM, Submit, SmallFA, GetVCF and Environment.



```
[jingjing@headnode PVCTools_alpha_v3.2]$ ./PVCTools
 SplitFA             - The source gene sequence FA file is split into small portions according to the included chromosomes.
 SegmentFA           - Each FA file is segmented into small portions according to the split number.
 SplitBAM            - The source ratio of the sample BAM file is split into small portions according to the included chromosome.
 SegmentBAM          - Each BAM file is segmented into small portions according to the split number.
 Submit              - Submit and calculate VCF results.
 SmallFA             - Calculate the small chromosome FA files.
 GetVCF              - Integration 'SplitFA' 'SegmentFA' 'SplitBAM' 'SegmentBAM' and 'Submit' steps.
 Environment         - Setting environment variable.
```

5.2 GetVCF module

This module can execute all the following modules in PVCTools, and give the final variation vcf file to users. The usage for this module is like:

./PVCTools GetVCF <-w WorkPath> <-fa FAPath> <-bam BAMPath> <-n SplitNumber> [-lm Size] [-R Reserved values] [-I] [-T Tools] [-P Tool'sParameters] [-q Queue] [-cpu CPU] [-span Span]

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -fa: The path of the original FA file.
  - -n: The number of divisions.
  - -bam: The path of the original comparison sample BAM files group.
- Optional Parameters:
  - -lm: Minimize the size(MB) of files to be cut(default: Cut all the FA files

imported from [falist]).

- -I: Will import [bamlist](By default, do not import, all of the BAM files in the target path will be calculated).
- -R: Add the reserved value at the end of the segment FA file (Default queue: 0).
- -T: The tool you want to use to run the task (Default tool: samtools). Optional tools: [samtools] [gatk] [freebayes].
- -P: Calculating tool's parameters.
- -q: The queue you want to run the task (Default queue: normal).
- -cpu: The number of CPUs you want to allocate for running the task (Default value: 1). Ideal value: The number of samples.BUT considering the CPU resources,this value should be less than the queue CPU maximum.
- -span: The maximum number of the CPU used on each node (Default value: 20).

5.3 SplitFA module

```
./PVCTools SplitFA <-w WorkPath> <-fa FAPath> [-q Queue] [-cpu CPU] [-span Span]
```

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -fa: The path of the original FA file.
- Optional Parameters:
  - -q: The queue you want to run the task (Default queue: normal).
  - -cpu: The number of CPUs you want to allocate for running the task (Default value: 1). Ideal value: The number of samples. BUT considering the CPU resources, this value should be less than the queue CPU maximum.
  - -span: The maximum number of the CPU used on each node (Default value: 20).

5.4 SegmentFA module

```
./PVCTools SegmentFA <-w WorkPath> <-n SplitNumber> [-lm Size] [-R Reserved values] [-q Queue] [-cpu CPU] [-span Span]
```

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -n: The number of divisions.
- Optional Parameters:
  - -lm: Minimize the size(MB) of files to be cut(default: Cut all the FA files imported from [falist]).
  - -R: Add the reserved value at the end of the segment FA file (Default queue: 0).
  - -q: The queue you want to run the task (Default queue: normal).
  - -cpu: The number of CPUs you want to allocate for running the task (Default value: 1). Ideal value: The number of samples. BUT considering the CPU resources, this value should be less than the queue CPU maximum.
  - -span: The maximum number of the CPU used on each node (Default value: 20).
- Tips: You may need to customize the FA file list that you want to import in [falist]. Running again will delete the previous data.

5.5 SplitBAM module

```
./PVCTools SplitBAM <-w WorkPath> <-bam BAMPath> [-l] [-q Queue] [-cpu CPU] [-span Span]
```

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -bam: The path of the original comparison sample BAM files group.
- Optional Parameters:
  - -l: Will import [bamlist](By default, do not import, all of the BAM files in

the target path will be calculated).

- -q: The queue you want to run the task (Default queue: normal).
- -cpu: The number of CPUs you want to allocate for running the task (Default value: 1). Ideal value: The number of samples. BUT considering the CPU resources, this value should be less than the queue CPU maximum.
- -span: The maximum number of the CPU used on each node (Default value: 20).

5.6 SegmentBAM module

./PVCTools SegmentBAM <-w WorkPath> <-n SplitNumber> [-R Reserved values] [-T Tools] [-q Queue] [-cpu CPU] [-span Span]

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -n: The number of divisions.
- Optional Parameters:
  - -R: Add the reserved value at the end of the segment FA file (Default queue: 0).
  - -T: The tool you want to use to run the task (Default tool: samtools). Optional tools: [samtools] [gatk] [freebayes].
  - -q: The queue you want to run the task (Default queue: normal).
  - -cpu: The number of CPUs you want to allocate for running the task (Default value: 1). Ideal value: The number of samples. BUT considering the CPU resources, this value should be less than the queue CPU maximum.
  - -span: The maximum number of the CPU used on each node (Default value: 20).
- Tips: You may need to customize the FA/BAM file list that you want to import in [falist]. Running again will delete the previous data.

5.7 Submit module

./PVCTools Submit <-w WorkPath> <-n SplitNumber> [-single] [-R Reserved values] [-T Tools] [-P Tool'sParameters] [-q Queue] [-single]

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
  - -n: The number of divisions.
- Optional Parameters:
  - -single: If you use this module separately, you have to use this parameter.
  - -R: Add the reserved value at the end of the segment FA file (Default queue: 0).
  - -T: The tool you want to use to run the task (Default tool: samtools). Optional tools: [samtools] [gatk] [freebayes].
  - -P: Calculating tool's parameters.
  - -q: The queue you want to run the task (Default queue: normal).
- Tips: You may need to customize the FA file list that you want to import in [falist].

5.8 SmallFA module

./PVCTools SmallFA <-w WorkPath> [-T Tools] [-P Tool'sParameters] [-q Queue]

- Required Parameters:
  - -w: Working directory path for using to store the generated files.
- Optional Parameters:
  - -T: The tool you want to use to run the task (Default tool: samtools). Optional tools: [samtools] [gatk] [freebayes].
  - -P: Calculating tool's parameters.
  - -q: The queue you want to run the task (Default queue: normal).

5.9 Environment module

./PVCTools Environment [-tool_name1 PATH] [-tool_name2 PATH] ...

- Optional Parameters:

- -tool_name: Setting tool's path. Optional tools: [samtools] [gatk] [freebayes] [bamUtil] [bcftools].

```
<PATH_SAMTOOLS>= /gpfs01/software/bio/samtools-1.3.1/samtools
<PATH_BCFTOOLS>=  /gpfs01/software/bio/bcftools-1.2/bcftools
<PATH_BAMUTIL>=/gpfs01/home/jingjing/software/bamUtil/bamUtil/bin/bam
<PATH_GATK>=   /gpfs01/home/jingjing/software/GATK/GenomeAnalysisTK.jar
<PATH_GATKCSD>=/gpfs01/home/jingjing/software/GATK/picard-tools-1.119/CreateSequenceDictionary.jar
<PATH_FREEBAYES>= /gpfs01/home/jingjing/software/freebayes/bin/freebayes
```

6. Examples

The test data contains 17 resequencing samples from Aarabidopsis. The bwa is used to align the raw data to Arabidopsis reference genome.

- Reference genome: TAIR10.fa
- Alignment files: align folder, which containing 17 bam files.

6.1 Setting environment

Command:

```
/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools Environment -samtools /gpfs01/software/bio/samtools-1.3.1/samtools -freebayes /gpfs01/home/jingjing/software/freebayes/bin/freebayes
```

6.2 Running in one command

6.2.1 Samtools

Command:

```
/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools GetVCF -w . –fa TAIR10.fa -n 100 -bam align/ -lm 10 -T samtools -q normal -cpu 50 -span 30
```

- -n: means divide the chromosome longer than **lm** into **n** pieces.

6.2.2 GATK

Command:

```
/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools GetVCF -w . –fa TAIR10.fa -n 100 -bam align/ -lm 10 -T gatk -q normal -cpu 50 -span 30
```

6.2.3 Freebayes

Command:

```
/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools GetVCF -w . –fa TAIR10.fa -n 100 -bam align/ -lm 10 –R 200 -T freebayes -q normal -cpu 50 -span 30
```

- Tips: -R: should be set longer than the length of read.

6.3 Running in separate command

- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SplitFA -w . -fa TAIR10.fa -q normal -cpu 20 -span 10`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SegmentFA -w . -n 30 -lm 10 -q normal -cpu 20 -span 10`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SplitBAM -w . -bam align/ -q normal -cpu 40 -span 30`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SegmentBAM -w . -n 30 -T gatk -q normal -cpu 40 -span 30`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools Submit -w . -n 30 -single -T gatk -q normal`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SmallFA -w . -T gatk -q normal`
- `/gpfs01/home/jingjing/project/tobacco/PVCTools/PVCTools_alpha_v3.3/./PVCTools SmallFA -w . -T gatk -q normal`