I have created dozens of small shell scripts to analyze SVs on a population scale. It is NOT easy for other people to read and use these scripts. So I will just give you overall ideas and principles.

1. Using “bwa sampe” to align reads to the reference genome in paired-end mode. Keep SAM results for the following analysis.

2. Using SVDetect to find SVs for each rice species. The input to the SVDetect program is “abnormal” read pairs.

SVDetect parameters

|  |  |
| --- | --- |
| mates\_orientation | FR |
| strand\_filtering | 1 |
| order\_filtering | 1 |
| insert\_size\_filtering | 1 |
| nb\_pairs\_threshold | 3 |
| nb\_pairs\_order\_threshold | 2 |
| indel\_sigma\_threshold | 6 |
| dup\_sigma\_threshold | 6 |
| singleton\_sigma\_threshold | 6 |
| final\_score\_threshold | 0.8 |

3. Using the filters developed by me to remove potential false positives from the SVDetect results. The programs and documentation can be found at <http://bioinfo.bti.cornell.edu/tool/SVFilter/>

|  |  |
| --- | --- |
| SV type | Filters applied |
| Deletion | Gap, ratio, SNV, read coverage |
| Insertion | Gap, ratio, SNV |
| Inversion | Gap, ratio, SNV |
| Tandem duplication | Sequencing depth (only 1 filter was used) |

Please pay attention that different SV types use different combinations of filters.

(1) gapfilter test\_sv genome.fa 1 -0.1

If parameter 3 is set to 1 and parameter 4 is set to a negative value, then the SV will be discarded as long as there is a gap within the SV region.

(2) ratiofilter test\_SV normPair.sam 0.2 75

75 is the read length depending on your data

(3) SNVfilter test\_sv genome.fa abnorm-pair.sam normPair.sam 2

(4) coveragefilter test\_sv genome.fa test.pileup -6 0.05 2

(5) depthfilter test\_sv chr-length test.pileup 1.5 1.5

4. Combine all SVs from all rice species, and clustered them into a tabular format.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | SV start | SV end | type | Species 1 | Species 2 |
| SV1 | Chr1 | 23123490 | 23128077 | DELETION | present | Absent |
| SV2 | Chr6 | 13191410 | 13196015 | DELETION | absent | Present |
| SV3 | Chr3 | 12277917 | 12283475 | DELETION | present | unknown |

5. Differentiate between “absent” and “unknown”. “Unknown” indicates that we are uncertain if an SV exists for a rice species because of low coverage, software limitation, etc.