1 Structural Variation Calling - Solutions

No questions in this section.

2 Looking at Structural Variants in VCF

2.1 Exercises

1. What does the CIPOS format tag indicate? **Confidence interval around POS for imprecise variants**
2. What does the PE tag indicate? **Number of paired-end reads supporting the variant across all samples**
3. What tag is used to describe an inversion event? **INV**
4. What tag is used to describe a duplication event? **DUP**
5. How many deletions were called in total? (**Hint:** DEL is the info field for a deletion. The -c option of the grep command can be used to return a count of matches.) **31, try**

```
grep -c "<DEL>" ERR1015121.vcf
```

6. What type of event is predicted at IV:437148? What is the length of the SV? How many paired-end reads and split-reads support this SV variant call? **Deletion -370 20 PE 21 split**

```
grep "437148" ERR1015121.vcf
```

7. What is the total number of SV calls predicted on the IV chromosome? **10, try**

```
grep -c "^IV" ERR1015121.vcf
```

3 Calling Structural Variants

Q: mean=454.87 std=86.29

3.1 Breakdancer

3.1.1 Exercises

```
grep "83065" ERR1015121.breakdancer.out
```

1. Inversion
2. -116,
3. 42

```
grep "258766" ERR1015121.breakdancer.out
```

4. Deletion (7325, 99)
5. `grep DEL | awk OFS= breakdancer.dels.bed | awk '{print $1"\t"$2"\t"$5"\t"$7"\t"$9}' > breakdancer.dels.bed`

3.2 Inspecting SVs with IGV

3.2.1 Exercises

1. Yes, a deletion (view as paired, sort by insert size, squish).
2. There are very few reads mapping, the reads that are mapped are of low mapQ and it has a SV score = 99
3. Size estimate? ~7.5k

Was the deletion at II:258766 also called by the other structural variant software and was the predicted size?

5. Yes, SVTYPE=DEL, SVLEN=-7438
6. DEL called by breakdancer (score=59). Not found by other caller Lumpy.
7. Yes, 2 reads support (red).

3.3 Lumpy

3.3.1 Exercises

1. The `-F` option in `samtools view` excludes reads matching the specified flag
2. reads in proper pair | read unmapped | mate unmapped | not primary alignment | PCR optical duplicate
3. Deletion -625
4. Deletion -369

4 Calling Structural Variants from Long Reads

4.0.1 Align the reads with minimap and convert to bam

```
minimap2 -t 2 -x map-pb -a ../ref/Saccharomyces_cerevisiae.R64-1-
1.dna.toplevel.fa.gz YPS128.filtered_subreads.10x.fastq.gz | samtools view -b
-o YPS128.filtered_subreads.10x.bam -
```

4.0.2 Sort the bam

```
samtools sort -T temp -o YPS128.filtered_subreads.10x.sorted.bam
YPS128.filtered_subreads.10x.bam
```

```
samtools calmd -b YPS128.filtered_subreads.10x.sorted.bam
../ref/Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz >
YPS128.filtered_subreads.10x.sorted.calmd.bam
```

4.0.3 Index the sorted bam

```
samtools index YPS128.filtered_subreads.10x.sorted.calmd.bam
```

4.0.4 Call SVs with sniffles

```
sniffles -m YPS128.filtered_subreads.10x.sorted.calmd.bam -v
YPS128.filtered_subreads.10x.vcf
```

4.0.5 Exercises

1. What sort of SV was called at on chromosome 'XV' at position 854271? __Deletion_
2. What is the length of the SV? **345**
3. How many reads are supporting the SV? **17 (RE tag)**
4. What sort of SV was called at on chromosome 'XI' at position 74608? __Insertion_
5. What is the length of the SV? **358**
6. How many reads are supporting the SV? **15**
7. How many inversions were called in the VCF? Note inversions are denoted by the type 'INV'. **None - no inversions were called**
8. How many duplications were called in the VCF? Note duplications are denoted by the type 'DUP'. **2**

5 Bedtools

5.1 Exercises

1. How many SVs found in `ERR1015069.dels.vcf` overlap with a gene? (**Hint:** Use bedtools intersect command) **18, try (note the -u parameter is required to get the unique number of SVs)**

```
bedtools intersect -u -a ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-1-
1.82.genes.gff3  | wc -l
```

2. How many SVs found in `ERR1015069.dels.vcf` do not overlap with a gene? (**Hint:** note the -v parameter to bedtools intersect) **9, try**

```
bedtools intersect -v -a ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-1-
1.82.genes.gff3  | wc -l
```

3. How many SVs found in `ERR1015069.dels.vcf` overlap with a more strict definition of 50%? **14, try**

```
bedtools intersect -u -f 0.5 -a ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-
1-1.82.genes.gff3  | wc -l
```

4. How many features does the deletion at VII:811446 overlap with? What type of genes? Note you will need to also use the -wb option in bedtools intersect. `bedtools intersect -wb -a ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-1-1.82.genes.gff3 | grep 811446` **4 features, all of them are protein coding genes (biotype=protein_coding)**
5. How many features does the deletion at XII:650823 overlap with? What type of genes? Note you will need to also use the -wb option in bedtools intersect. `bedtools intersect -wb -a`

`ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-1-1.82.genes.gff3 | grep 811446` **2 features, all of them are protein coding genes (biotype=protein_coding)**

6. What is the closest gene to the structural variant at IV:384220 in `ERR1015069.dels.vcf`? **YDL037C, try**

```
bedtools closest -d -a ERR1015069.dels.vcf -b Saccharomyces_cerevisiae.R64-1-1.82.genes.gff3| grep IV | grep 384220
```

5. How many SVs overlap between the two files `ERR1015069.dels.vcf` and `ERR1015121.dels.vcf`? **27, try**

```
bedtools intersect -u -a ERR1015069.dels.vcf -b ERR1015121.dels.vcf | wc -l
```

6. How many SVs have a 90% reciprocal overlap between the two files `ERR1015069.dels.vcf` and `ERR1015121.dels.vcf` (**Hint:** first find the option for reciprocal overlap by typing: bedtools intersect -h) **24, try**

```
bedtools intersect -u -r -f 0.9 -a ERR1015069.dels.vcf -b ERR1015121.dels.vcf | wc -l
```