1 Read Alignment

There are no questions in this section.

2 Performing Read Alignment

1. 145,441,459

2. The sam is ~ 157M and the bam is ~ 25M

3. 67,461 (17.2%). This is the number of unpaired read dups (359) + the number of paired read dups (33551 * 2)

4. 392,820

5. 391,603/392,820 (99.7%)

6. 389,410

7. 0

8. 419 (mean) 113.9 (standard deviation)

9. 7,853 (2.0%)

10. First/Forward read

3 Alignment Visualisation

1. This exercise is just asking you to explore the genome and become familiar with navigating in IGV.

2. 23X-57X

3.

There is a 1bp insertion (at "T)" at position 87,483,966. This is supported by 9 reads.

There is a 28bp deletion at position 87,483,966. This is supported by 3 reads.

The third mutation is a bit harder to spot, because it's bigger than the read length (so no single read will span it). Look first at the coverage track, and notice a sharp coverage drop between chr7:87,483,833 and chr7:87,484,169. This suggests that one of the two alleles has been deleted across that position. Notice also the soft-clipping of some reads "entering" chr7:87,483,833 from the left, and the same softclipping of reads entering chr7:87,484,169 from the right. This soft-clipping shows up as reads being partly multi-coloured. That's happening because the physical genome between those points has been excised for one allele, causing the mis-alignment when we attempt to align some reads to the reference genome. That mis-alignment causes the bwa aligner to "give up" and mark a part of the read as soft-clipped.

4.

This mouse was bred from a zygote which was mutagenised with Crispr-Cas9, targeted at the Tyr locus. You are watching the zygote DNA-repair machinery panicking and grabbing at straws when trying to repair double-stranded DNA breaks. In the process, it makes mistakes, and those mistakes are propagated into the mouse genome: different zygote cells received different mutations, which is why some reads reflect different mutations to others. Specifically - The CRISPR-Cas9 has acted on the zygote at this locus to create Non-Homologous-End-Join-based damage around 87,483,960: that resulted in a subclonal 1bp insertion and a 28bp deletion. Also, a related DNA repair process has resulted in the a 336bp deletion across the same area.

3.1 Alignment Workflows

1. -M marks shorter split hits as secondary and -R adds the read group to the header of the BAM file

2. -b means create a BAM as output and -S indicates that the input file is a SAM file. The -S option is now ignored by samtools as it can autodetect the input file type.

3. 397,506

4. 303,036/397,506 (76.2%)

5. 282,478

6. 2,239

7. 275.9 (mean) and 47.7 (standard deviation)

8. 23,789 (7.9%)

9. First

10.

```
bwa mem -M -R "@RG\tID:lane2\tSM:60A_Sc_DBVPG6044" ../../../../ref/Saccharomyces_cerevisiae.R64
1.dna.toplevel.fa.gz s_7_1.fastq.gz s_7_2.fastq.gz | samtools view -bS - |
samtools sort -T temp -O bam -o lane2.sorted.bam -
```

~22M

Merge: `~/course_data/read_alignment/data/Exercise2/60A_Sc_DBVPG6044/library1$ picard MergeSamFiles -I lane1/lane1.sorted.bam -I lane2/lane2.sorted.bam -O library1.bam`

Markdup: `~/course_data/read_alignment/data/Exercise2/60A_Sc_DBVPG6044/library1$ picard MarkDuplicates -I library1.bam -O library1.markdup.bam -M library1.metrics.txt`

11. 12399 or 3115 + (4642 * 2) = unpaired read dups + (paired read dups *2)

12. 2.5%

3.2 Exercises

1. No answer needed

2. The reference base is C

3. No (the reads call T)

4. The reference base is G and all reads agree

5. No answer

6. An insertion

7. No answer

8. A deletion. This is unlikely to be a true variant and may be due to misalignment due the run of T's in the flanking region.

9. The following command produces a CRAM file which should be ~29MB in size.

```
samtools view -C -T ../../../ref/Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa
-o library1.markdup.cram library1.markdup.bam
```

-C means create a CRAM file as output

-T is the reference file to use for the compression

-o is the name of CRAM file to create