**Read Alignment**

1. **How many reads were removed?**

Removed = Input (50,000 reads (25k pairs × 2) – reads that passed filter (48,854) = 1,146 reads (1,120 low-quality + 26 with too many Ns).

1. **What percentage of bases are Q30? Do you think the reads are high quality based on this?**

R1 after filtering: 96.7% Q30 and R2 after filtering: 94.0% Q30. Yes, these are very high-quality reads as they are above 85–90% Q30 and can be considered as good Illumina sequencing data.

1. **How many reads are duplicated?**

Duplication rate is 4.784% of 48,854 reads passing filters. Therefore, duplicated reads = 2,337.18 reads.

1. **Is there a clear decline in quality for Read 1 and Read 2, if so, why do you think that is?**

Yes. Read 2 has lower quality than Read 1. It may be due to as sequencing progresses, since Read 2 is sequenced after Read 1 (using the same DNA clusters), its overall quality tends to be lower or due to phasing errors.

**Running bowtie2**

1. **how long did this take to run? How long do you think it would take to align 10 million reads?**

It took about 10 -15 seconds to run and at the same rate, aligning one million reads might take 2 hours.

1. **What is the overall read alignment rate for both samples? which one is higher? Do you think this is a good alignemnt rate?**

It is 94% with sample 1 being slightly higher. I believe it as a good alignment rate, as most reads mapped successfully to the reference genome.

1. **What could the reads that don't map to the genome represent?**

They could represent possible technical or signal errors, contamination during sequencing by adapters, etc.

1. **Load the first few lines of one of the .sam files. Where are the first three reads mapping to the genome? Is it a perfect/exact map?**

The first few reads from the .sam file, map to the chromosome 10A that is around 23 Mb, chromosome 10B that is around 20 Mb and chromosome 1A that is around 40 Mb. I do not think that they are perfectly aligned.

1. **What is the average coverage of this data?**

It is about 17X in this region.

1. **How many variants do you think there are? What about variants where there is a difference in 1 read, but the others match the reference genome?**

It looks like there are around 8 positions that differ from the grey regions, which means the same number of variants. Variants that appear in only one read while all others match the reference genome, might be the result of sequencing errors that could include errors due to machine noise/technical issues, sequencing chemistry or other factors, like genetical divergence could also play a role.