Removing duplicates from alignment output

If you see in your bam file, that reads are piling up with same start end end coordinates, these may be pcr duplicates, which should be removed/flagged in your bam files.

Picard MarkDuplicates is the preferred tool for this, but it is very fickle with the type of bam file it will work on.

Samtools can be an easier option to start with for removing potential pcr duplicates in your data.

1. (OPTIONAL) samtools fixmate

Because samtools rmdup works better when the insert size is set correctly, samtools fixmate can be run to fill in mate coordinates, ISIZE and mate related flags from a name-sorted alignment.

```
samtools fixmate <in.nameSrt.bam> <out.bam>
```

2. samtools rmdup -sS <input.srt.bam> <out.bam>
Remove potential PCR duplicates: if multiple read pairs have identical external coordinates, only retain the pair with highest mapping quality. In the paired-end mode, this command ONLY works with FR orientation and requires ISIZE is correctly set. It does not work for unpaired reads (e.g. two ends mapped to different chromosomes or orphan reads).

OPTIONS:
-s Remove duplicate for single-end reads. By default, the command works for paired-end reads only.
-S Treat paired-end reads as single-end reads.

```
default:
samtools rmdup <input.bam> <output.bam>
or
samtools rmdup -s <input.bam> <output.bam>
```

Load the output.bam file into IGV to check on areas which showed evidence of pcr duplicates before.