Finding structural variations in a genome after comparison with a reference genome

In the directory you will find the genomic reference sequence of a bacterium called *Lactobacillus casei*. The genome is 3,079,196 bp long. The aim of the project is the "resequencing" of a similar bacterium that we call *lact.sp*, using mate pairs reads. In particular we want to see if *lact.sp* has genomic structural variations. We suspect that the *lact.sp* genome may have small and large deletions, insertions and inversions, and we want to find where they are. The mate pairs fastq reads are supplied in two files.

You should do the following tasks:

Part 1 (see unix.pdf file for more details)

1) Download the reference genome and the zip file with the fastq reads
2) Install BWA on your computer
3) Use BWA to align the illumina reads on the reference genome
4) Install IGV on your computer
5) Process the sam file obtained by the BWA analysis to make a bam file
6) Sort and index the bam file (see unix.pdf)
7) Visualize the coverage and mate pairs on the IGV
8) Manually find and comment any “anomalous” pair of mates

Part 2 (see wig.pdf and physical.pdf for more details)

9) Implement in the language of your choice a program to create a wig file with physical covarage. An example can be found in physical.pl, but it would be nice to implement the suggestions described in physical_coverage.pdf.
10) Create a wig track with sequence coverage and compare it with the IGV track
11) Read the sam file and for each mate pair calculate the length of the genomic insert; then calculate the mean and standard deviation of the inserts, possibly discarding those that are totally out of range. A plot of the length distribution may help to evaluate the sizes.
12) Create a track with the percentage of inserts with a length exceeding n standard deviations (for instance n=2) above or below the mean.
13) Comment the results

Part 3

Be creative and make other tracks to identify other peculiarities associated to structural variations. Here there are some ideas that may help:

14) Create a track with the covarage of "unique" reads. Unique reads are those mapping on a single genomic position. See the “flag” specifications of the sam file.
15) Create a track with the covarage of "multiple" reads. The multiple reads are those mapping in more than one genomic position.
16) Create a track with the percentage of oriented mates. The orientation can be found on "bit 16" of the "flag" (see sam specifications). What could they indicate?
17) Create a track with percentage of single mates (see sam specifications). Single mates are those where only one of the two reads is mapping. It would be even better if the peack is done in correspondence of the missing read. What could this track indicate?
18) Create a track with the average length of the physical inserts covering the position of the genome. Compare the results with those from point 12.
19) Create tracks with the kmer frequence.
20) Create a track reporting the presence of “H” or “S” in the CIGAR field. What could they indicate?
21) Create a track with the gene predictions, using positive and negative values to show genes on the forward and reverse DNA strand. For this task the start and stop codons should be considered and the all the "open reading frames" longer than 300 bases should be displayed.