HIV analysis

The full analysis is consisted of many steps and we will here to through them slowly

The alignments are stored in sam/bam (see samtools specification for more details). The difference between sam and bam is that the former is stored in plain text and the latter is compressed

An (simple) outline of the alignment, consensus sequence and variant calling process:

1. Trim away sequencing adaptors
2. Align reads to the reference with SMALT . A lot of other aligners are available such as BWA and bowtie2
3. Sort the aligned reads according to chromosome and position using samtools
4. Remove/mask read pairs that occur more than once (duplicate reads) with picard
5. fix indel mis-alignments
6. Index the final alignment file (bam)
7. Do various analyses on the final bam file:
    a. consensus calling
    b. codon usage
    c. QC

```
#make a local directory and copy the test dataset into it
mkdir HIV_analysis
cp /software/packages/training/HIV/* HIV_analysis
cd HIV_analysis

# The directory should contain 2 files the fastq files (hiv_raw.1.fq.gz,
hiv_raw.2.fq.gz)

# remove the sequencing adaptors
/software/packages/cutadapt-1.1/bin/cutadapt -b TGTAGAACCATGTCGTCAGTGT -b
AGACCAAGTCTCTGCTACCGT hiv_raw.1.fq | gzip -c > hiv.1.fq.gz
/software/packages/cutadapt-1.1/bin/cutadapt -b TGTAGAACCATGTCGTCAGTGT -b
AGACCAAGTCTCTGCTACCGT hiv_raw.2.fq | gzip -c > hiv.2.fq.gz


# align the reads to the reference (run smalt-0.7.6 to see all options)
/software/bin/smalt-0.7.6 map  -f samsoft /refs/HIV/K03455_s1k6 hiv.1.fq.gz
hiv.2.fq.gz  > HIV.sam

# look at the sam file to ensure that it is correct and compare it with the samtools
specification (link above) to understand the file format
less HIV.sam

# Make the samfile into a bam file using samtools
/software/bin//samtools view -Sb HIV.sam -o HIV.bam

# sort the bamfile with samtools
/software/bin//samtools sort HIV.bam HIV_sorted



# Mark/remove duplicate reads to ensure better consensus calling:
/software/bin//picard -T MarkDuplicates  I= HIV_sorted.bam O= HIV_rmdups.bam AS=true
M=rmdup.csv
```

```
# index the bam file so it can be viewed in IGV later on
/software/bin//samtools index HIV_rmdups.bam

# Look at how well the mapping was done by looking at the flagstats from samtools
/software/bin/samtools flagstat HIV_rmdups.bam

# Smalt sometimes does odd alignments, so fix these:
/software/packages/ctru-clinical/scripts/bam_fix_indels.pl HIV_rmdups.bam
HIV_fixed.bam

# index the bam file so it can be viewed in IGV later on
/software/bin//samtools index HIV_fixed.bam



#
_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-
=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_-=-_
# primary analysis is all done, now it is time to do the QC & secondary analysis

# First overall coverage QC
/software/packages/ctru-clinical/scripts/HIV_sample_QC.py HIV_fixed.bam
# This program is going to change soon, so if this does not work try:
/software/packages/ctru-clinical/scripts/VIRUS_sample_QC.py HIV_fixed.bam


# Then calculate the consensus from the various regions:
/software/packages/ctru-clinical_dev/scripts/Bam_consensus.py HIV_fixed.bam 2253 4227
>> HIV_consensus.txt
/software/packages/ctru-clinical_dev/scripts/Bam_consensus.py HIV_fixed.bam 4232 5099
>> HIV_consensus.txt

# And finally the codon usage for the <20% detection level
/software/packages/ctru-clinical/scripts/codon_usage.py HIV_fixed.bam >
HIV_fixed_codons.xls
```

```
# Now load up the two bamfiles in IGV and see what the raw data looks like
# and look at the data in all those nice files you have created.
```