Derived allele frequencies, annotation liftover

Last time we made a genomic alignment for nine-spined and three-spined sticklebacks and used that to (1) infer ancestral alleles for our binary SNP variants, and (2) create a lift-over chain that can be used to transfer genomic coordinates between the species. Now we use those in practice.

DAF

```
cd ~/session_7
mkdir daf
```

Get the sample names for a few populations:

```
bcftools query -l data101_goodAA.vcf.gz | grep popA > daf/popA.txt
```

Do the same for popB, popC, popG, popF and popJ.

Use then `vcftools` to compute derived allele counts (see http://vcftools.sourceforge.net/man_latest.html for details):

```
bcftools view -S daf/popA.txt -e 'INFO/AA="."' -m2 -M2 data101_goodAA.vcf.gz |
vcftools --vcf - --counts --derived --out daf/popA_derived
```

The output of that is not easy to parse within R so we do it with a bit of awk code:

```
less -S daf/popA_derived.frq.count

awk '{if($4==20){print substr($5,3),substr($6,3)}}' daf/popA_derived.frq.count |
sort -n -k2 | uniq -c > daf/popA_derived.txt

less -S daf/popA_derived.txt
```

This is now easier. Do the same for popB, popC, popG, popF and popJ.

Use R to plot the data:

```
popA = read.table("daf/popA_derived.txt",header=F)
popB = read.table("daf/popB_derived.txt",header=F)
popC = read.table("daf/popC_derived.txt",header=F)
popG = read.table("daf/popG_derived.txt",header=F)
popF = read.table("daf/popF_derived.txt",header=F)
popJ = read.table("daf/popJ_derived.txt",header=F)

colors=rainbow(6)

#png("daf/daf_six.png",800,600)
par(lend=1,mfrow=c(1,1),lwd=2,pch=1)
```

```
plot(1:19,popA[2:20,1]/sum(popA[2:20,1]),type="b",xlab="derived alleles
(N=20)",ylab="prop. SNPs",ylim=c(0,0.35),xaxt="n",col=colors[1],pch=0)

lines(1:19,popB[2:20,1]/sum(popB[2:20,1]),type="b",col=colors[2],pch=1)
lines(1:19,popC[2:20,1]/sum(popC[2:20,1]),type="b",col=colors[3],pch=2)
lines(1:19,popG[2:20,1]/sum(popG[2:20,1]),type="b",col=colors[4],pch=3)
lines(1:19,popF[2:20,1]/sum(popF[2:20,1]),type="b",col=colors[5],pch=4)
lines(1:19,popJ[2:20,1]/sum(popJ[2:20,1]),type="b",col=colors[6],pch=4)

axis(1,1:19,1:19,cex.axis=0.75)

legend("topright",legend=c("popA","popB","popC","popG","popF","popJ"),col=colors,bt
y="n",ncol=6,pch=c(0:9))

#dev.off()
```

Compare the DAF plot to the IBS clustering tree. What do you notice? Redo the filtering (`select1` and `select2`) from previous time with the vcf file containing the AA's. Redo the DAF plots. Do you see differences?

Annotation liftover

Ensembl provides lots of raw data through their ftp site. You can start from here: http://ftp.ensembl.org/pub
We want the latest version ("current") of gene annotations in gff3 format. We therefore go to folder "current_gff3" and find there the three-spined stickleback "gasterosteus_aculeatus". If we know the address, we can get it directly with wget:

```
wget http://ftp.ensembl.org/pub/current_gff3/gasterosteus_aculeatus/\
Gasterosteus_aculeatus.BROADS1.84.gff3.gz -P reference
```

Look into the file and go down until you find a line not starting with a hash:

```
less -S reference/Gasterosteus_aculeatus.BROADS1.84.gff3.gz
```

This file contains the genomic coordinates of known genes in three-spined stickleback. We want to transfer those coordinates to the genome of nine-spined stickleback. This transfer is based on the assumption that the two species have the same genes in homologous genomc locations and that similar sequence regions detected in genomic alignment are homologous. Kent utilities contain program "liftOver" to transfer bed files but a new program called CrossMap (http://crossmap.sourceforge.net) is more flexible and can transfer also gff and vcf files. See the options of the program:

```
CrossMap.py

CrossMap.py gff
```

and build then a command to do the iftover:

```
CrossMap.py gff reference/threeToNine.liftOver.gz \
reference/Gasterosteus_aculeatus.BROADS1.84.gff3.gz reference/ninespine.gff3
```

If we look at the transferred annotations:

```
less -S reference/ninespine.gff3
```

one of the first things we notice is that contig "deg7180000006464" appears to contain mitochondrial genes ("mt_gene"). We can look into that contig using "samtools tview":

```
samtools tview ../session_3/sample-1_realn.bam reference/ninespine.fa \
-p deg7180000006464
```

Do you agree with that? Why?

We have earlier imported a gff file containing repeat annotations to IGV. Open now the vcf file in IGV and import also the newly generated gene annotation.

Alternative to IGV

Those wanting to experiment may test a new program called BasePlayer for analysis and visualisation of genomic data. It is developed by Riku Katainen (from CS, now in Meilahti with Lauri Aaltonen) and is found at https://www.cs.helsinki.fi/u/rkataine/BasePlayer/. The official version comes with the human genome and the package takes over 3GB of space. A smaller version (without human genomes) is available here: http://wasabiapp.org/vbox/data/session_8/BasePlayer.tgz

New genomes can be added by copying the Fasta and Fasta.fai files to the folder "genomes". Gff3 files can be opened as tracks. The files have to compressed (bgzip) and indexed (tabix) bed files. The one that I created for nine-spined stickleback is broken (you can try to make a valid one!) and is not shown. The program can be used to visualise vcf and bam data, though.