hw2

Maria Firulyova

```
library(data.table)
library(ggplot2)
library(dplyr)
library(tidyr)
```

1

Assume you study a particular group of individuals with two hypothetic subpopulations of equal size with the following allele frequencies in a biallelic locus:

subpopulation 1) p = 0.9, q = 0.1

subpopulation 2) p = 0.6, q = 0.4

Fst = (Ht - Hw) / Ht

```
fst <- function(p1, q1, p2, q2) {
  pt <- ((p1 + p2) / 2)
  qt <- ((q1 + q2) / 2)
  hw <- ((2 * p1 * q1) + (2 * p2 * q2)) / 2
  ht <- 2 * pt * qt
  (ht - hw) / ht
}

fst(p1=0.9, q1=0.1, p2=0.6, q2=0.4)
```

```
## [1] 0.12
```

Answer: Fst = 0.12

2

NUMINIDS:: 45

EXTRACOLS: 1

NUMLOCI: 14

1) Run structure with different values of clusters (-K option) - 3, 4, and 5.

```
clusters = range(3, 6)

rule all:
    input: expand("out/{k}_out.txt", k=clusters)

rule structure:
    output: out="out/{k}_out.txt"
    params: k=lambda wildcards: wildcards.k, def_output='outfile_f'
    log: "logs/{k}.txt"
```

```
shell:
    """
    ./structure -K {params.k} 2> {log}
    mv {params.def_output} {output.out}
    """
```

- Run program:

```
snakemake -j 1
```
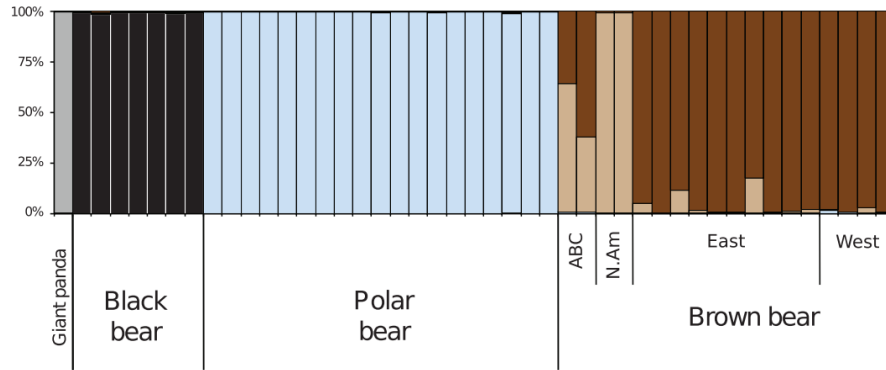
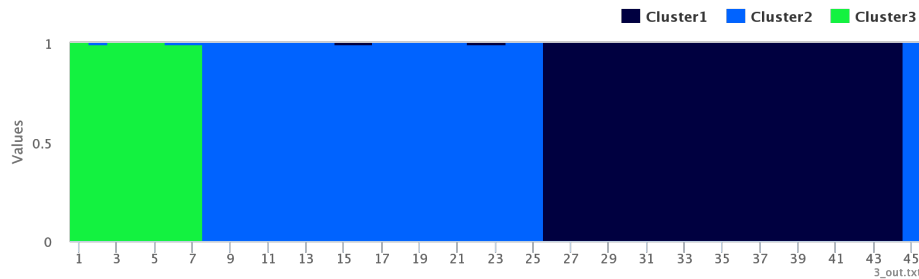- Visualize the results using Pophelper (http://pophelper.com/) or R.

Figure 1: K = 3

Figure 2: K = 3

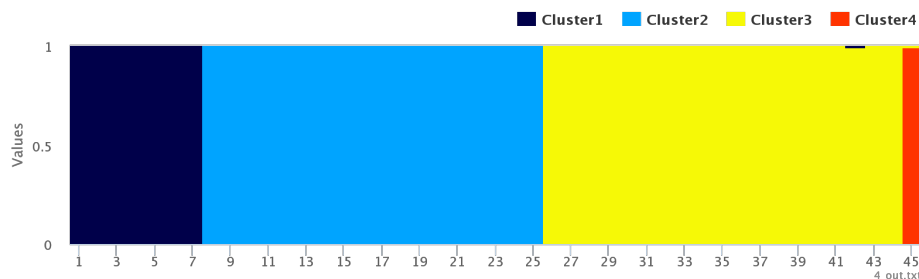
Figure 3: K = 4

- Which of the plots (for K = 3, 4, or 5) corresponds to the Figure 2 of the original paper?

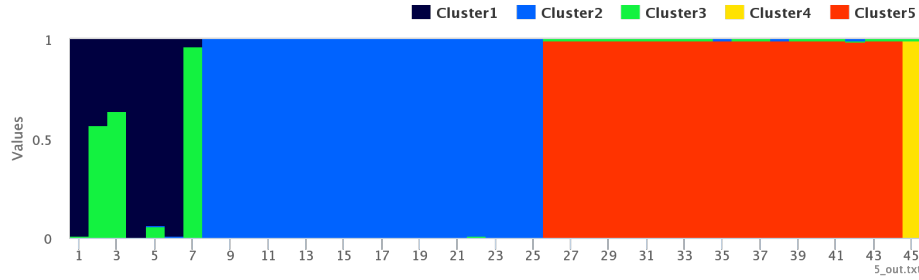Plot which build using `K = 5` parameter corresponds to the Figure 2 of the original paper.

Figure 4: K = 5

2) Do the data obtained prove that the white bears are distinct from brown and black ones?

The data obtained by the authors are more relevant than past studies through the use of different nuclear loci, since comparison solely for mitochondrial DNA can introduce certain bias in the results (in particular, because mtDNA is inherited maternally). Studies based on the analysis of genome-wide data have shown that the flow of genes between different species of bears, including between white and brown bears, occurred in the process of evolution. Since gene flow left its mark in different regions of the genome, perhaps the researchers whose article we read should have taken more loci for comparison: at least this would increase the accuracy of the date of divergence of white and brown bears.

3

Let's assume you study a sample of 4 alleles from a particular population. You've made a phylogenetic analysis of these sequences and got the following genealogy together with the coalescence times (t1, t2, t3):
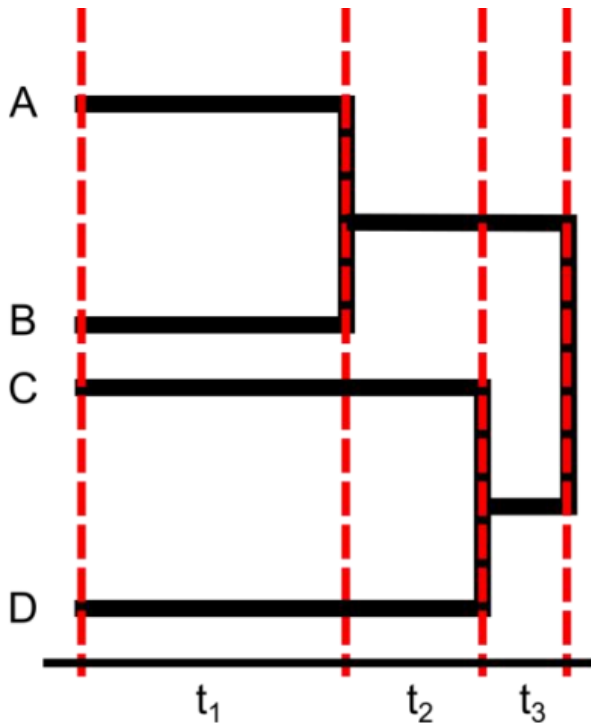

Figure 5: Third task

Let t1 = 6000, t2 = 5000, t3 = 3000. Use the logic of the classical skyline method. What is the demographic history of the population? Was the population size increasing or decreasing recently?

```
classic_sky <- function(k, t) {
  0.5 * (k * (k - 1)) / 2 *  t
}
```
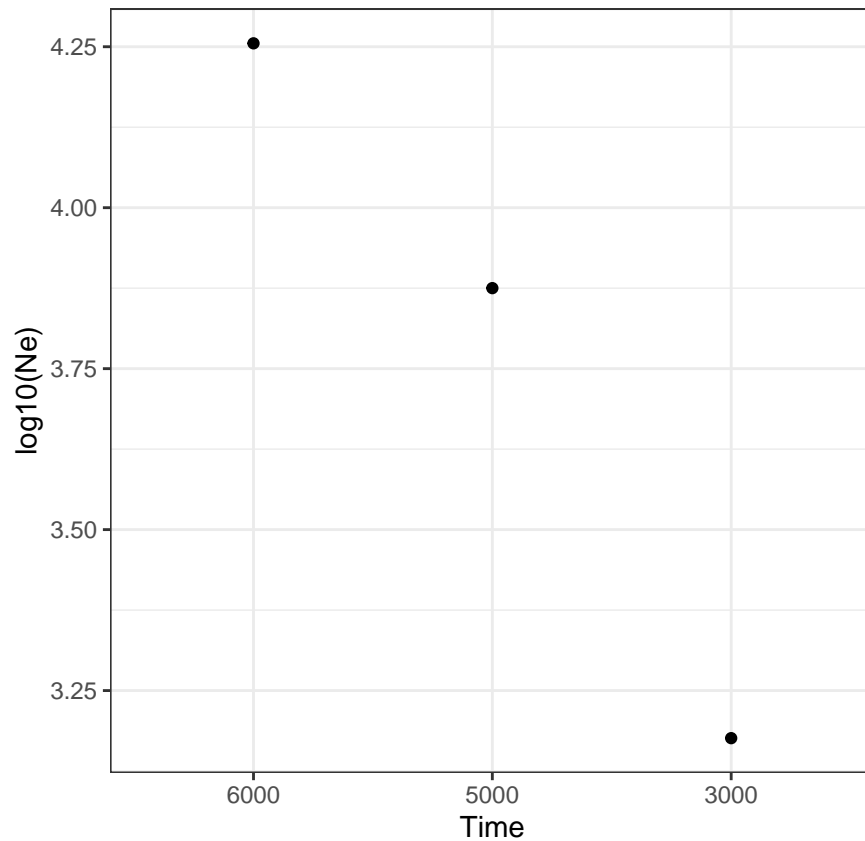
```
ne4 <- classic_sky(4, 6e3)
ne3 <- classic_sky(3, 5e3)
ne2 <- classic_sky(2, 3e3)
```

Let's calculate Ne for each case:

- $\text{Ne}(4) = 1.8 \times 10^4$

- $\text{Ne}(3) = 7500$

- $\text{Ne}(2) = 1500$

```
df <- cbind(c(ne4, ne3, ne2), c(6e3, 5e3, 3e3)) %>%
  as.data.frame() %>%
  magrittr::set_colnames(c('Ne', 'Time')) %>%
  mutate(Time = factor(Time, levels = c(6e3, 5e3, 3e3)))

ggplot(df, aes(x = Time, y = log10(Ne)))+
  geom_point()+
  theme_bw()+
  theme(aspect.ratio = 1)
```


The population size has decreased across time.

4

Let's make some Bayesian skyline analysis to model demographic histories. We'll use the data from Atkinson et al., 2007. We'll need BEAST2 tool and the Tracer application to run the analysis and view the results.

.xml : configuration file built using BEAUTi.

Questions:

1) What is the (approximately) median log of posterior probability as shown by Tracer? Please provide the plot of the distribution.

2) How does the population size change over time? Were there any bottlenecks? Please provide the skyline plot.

3) Compare your results with plots from the original paper. Which population was used in your analysis? Are there any differences in Ne values?