F statistics

Shyam Gopalakrishnan

Aug 8th 2019

Overview

- Comparing two populations
    - Trees to show population relationships
    - F statistics
    - D statistic
- Genetic affinities using F statistic
- Admixture signals using D statistics

How to compare two groups

Group 1: -0.2609 0.3309 0.6095 -0.8065 -0.6407 -0.4874 0.8393 -0.343 1.0843 0.4343

Group 2: -0.2732 -1.0104 1.0404 -0.1881 -1.1271 -0.2672 -0.5584 0.7958 -1.3073 1.1832

What methods would you use to compare these two sets of numbers?

-- How similar are they?
-- How different are they?

Comparing two populations

- Using similar ideas as the previous slide, how would you compare two populations?

Comparing two populations

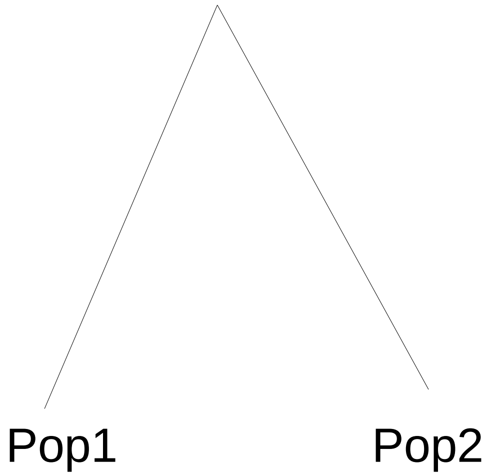
- Using similar ideas as the previous slide, how would you compare two populations, given genetic information on them?

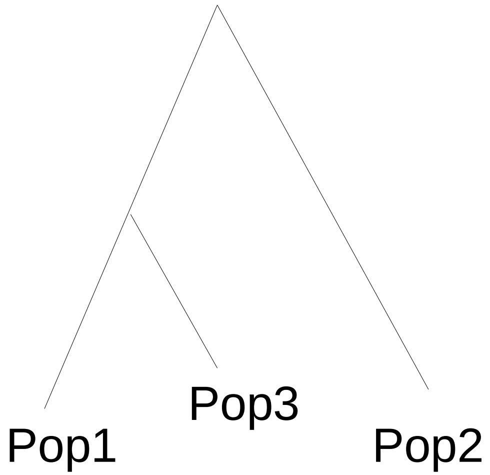
Comparing two populations

- Using similar ideas as the previous slide, how would you compare two populations, given genetic information on them?
    - Use allele frequencies
        - Correlations
        - Differences

Using trees to relate populations

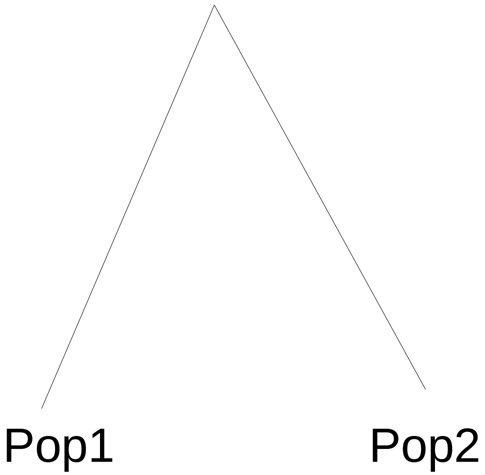
- Trees are often used to show the relationships between populations.

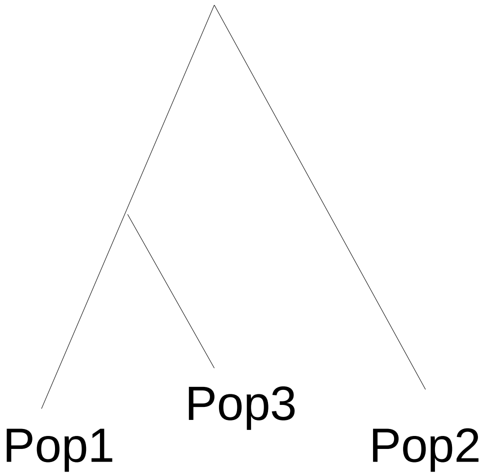




Measuring distances between populations on a tree?

- How far is pop1 from pop2, in the first tree?
- How far is ancestor of pop1 and pop3 from pop2





What do these branch lengths on trees mean?

- Branch lengths are often related to genetic drift
  - Genetic drift can be thought of as the rate at which frequencies change, under neutrality
  - Drift depends on demography – most notably, on population size

F statistics

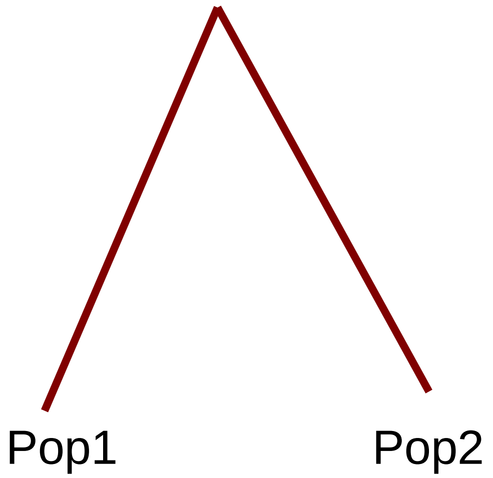
- Let us first define an F statistic
  - Given four populations A, B, C, D with allele frequencies a, b, c and d respectively,

    F(A,B; C,D) = **E[**(a-c)(b-d)**]**
  - Here think of **E[]** as average over all SNPs.

F2 statistic or pairwise distance

- If A == B and C == D, then we reduce to the two population F2 statistic.

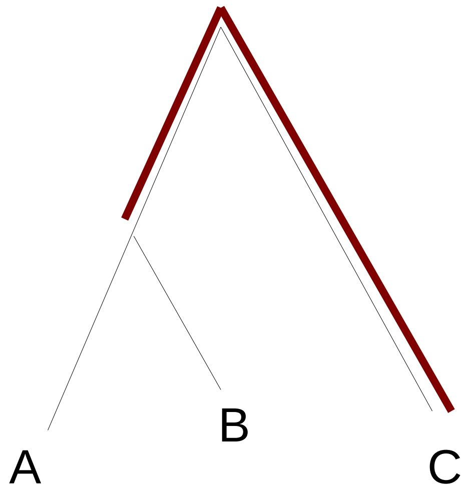
$$F2(A,B) = \mathbf{E[}(a-b)^2\mathbf{]}$$



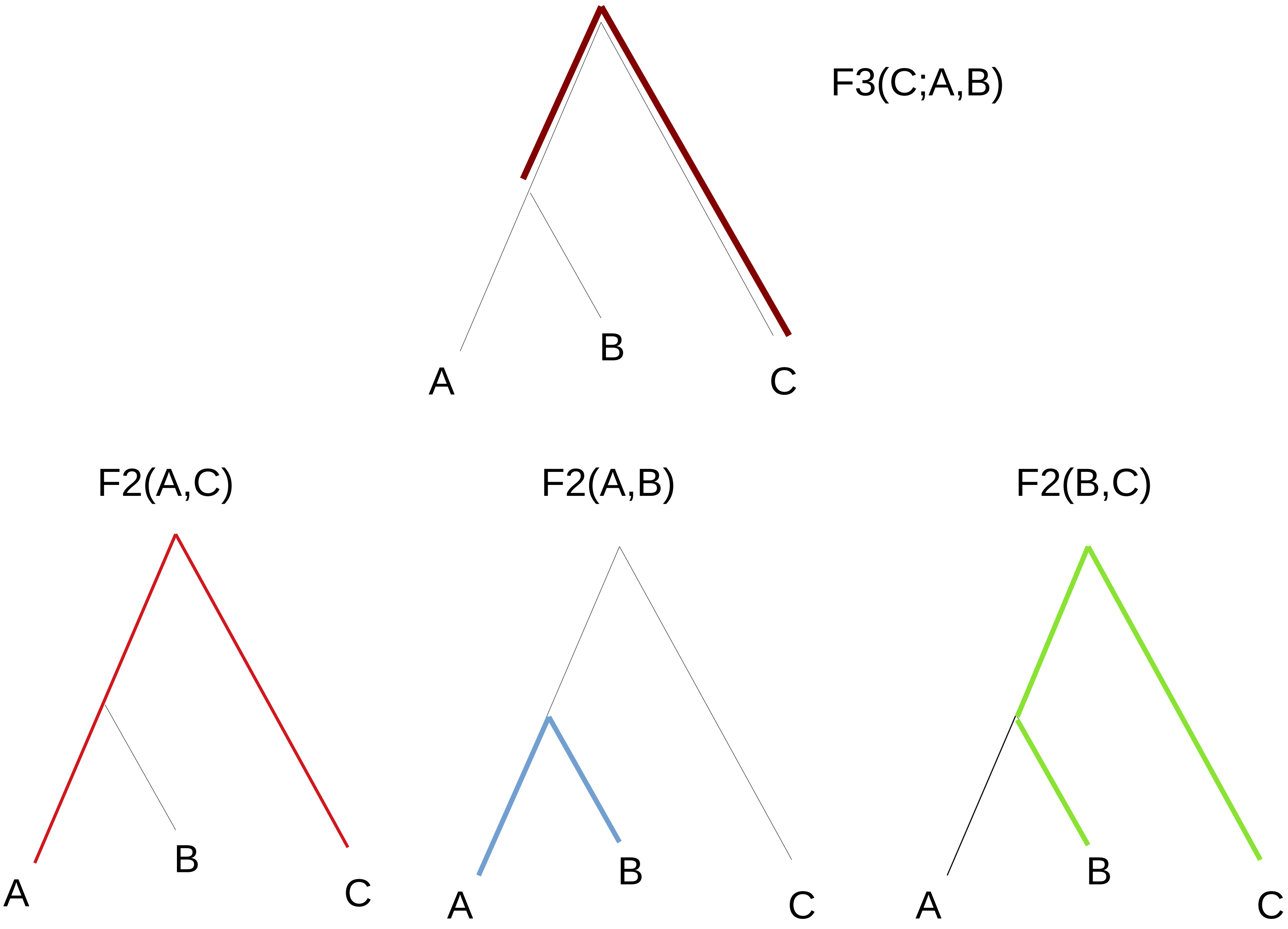
F3 statistic

- With three populations, one gets the F3 statistic.

F3(C;A,B) = **E[**(a-c)(b-c)**]**



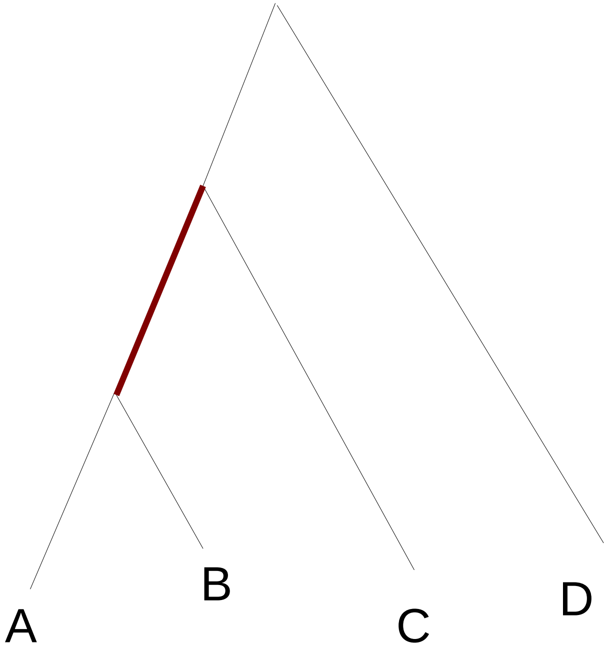
F3 stat in terms of F2 stat



F4 statistic

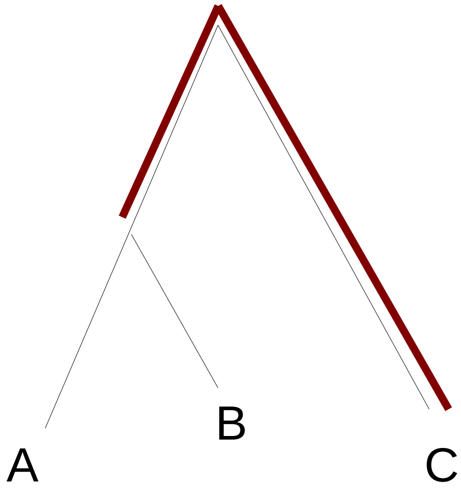
- Four distinct populations

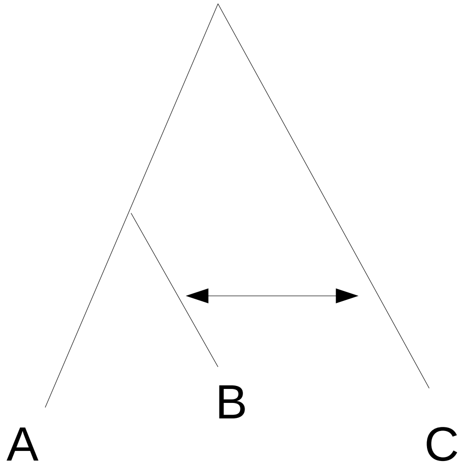
$$F4(A,B;C,D) = \mathbf{E[}(a-c)(b-d)\mathbf{]}$$



Using F-statistics

- Testing for treeness with F3
  - Admixture F3
  - Think back to getting F3 from F2 stats.





Using F-statistics

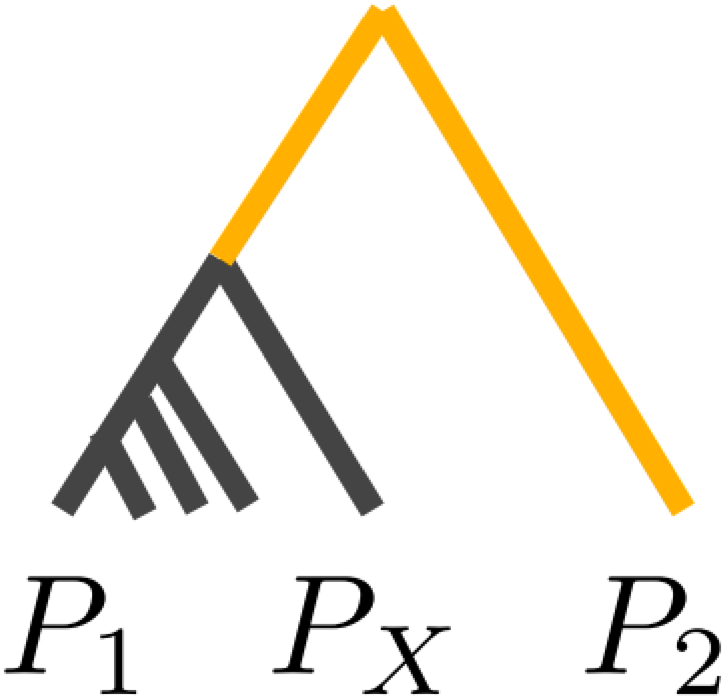
- Suppose you have an unknown population, P1, and you want to figure out which population it is genetically closet to. How would you do it?

Using F-statistics

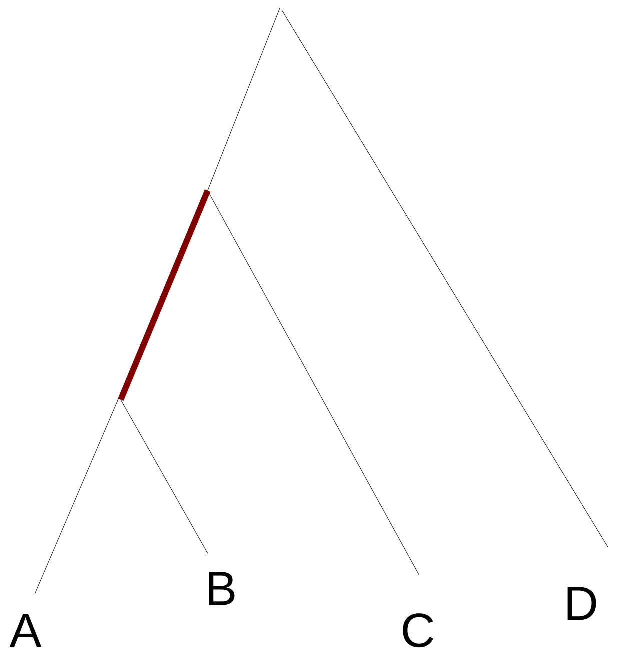
- Suppose you have an unknown population, P1, and you want to figure out which population it is genetically closet to. How would you do it?
  - Outgroup F3 statistic

Outgroup F3 statistic

- Why outgroup F3?
  - Individual population
    do not matter
  - Sampling times of PX
    matter



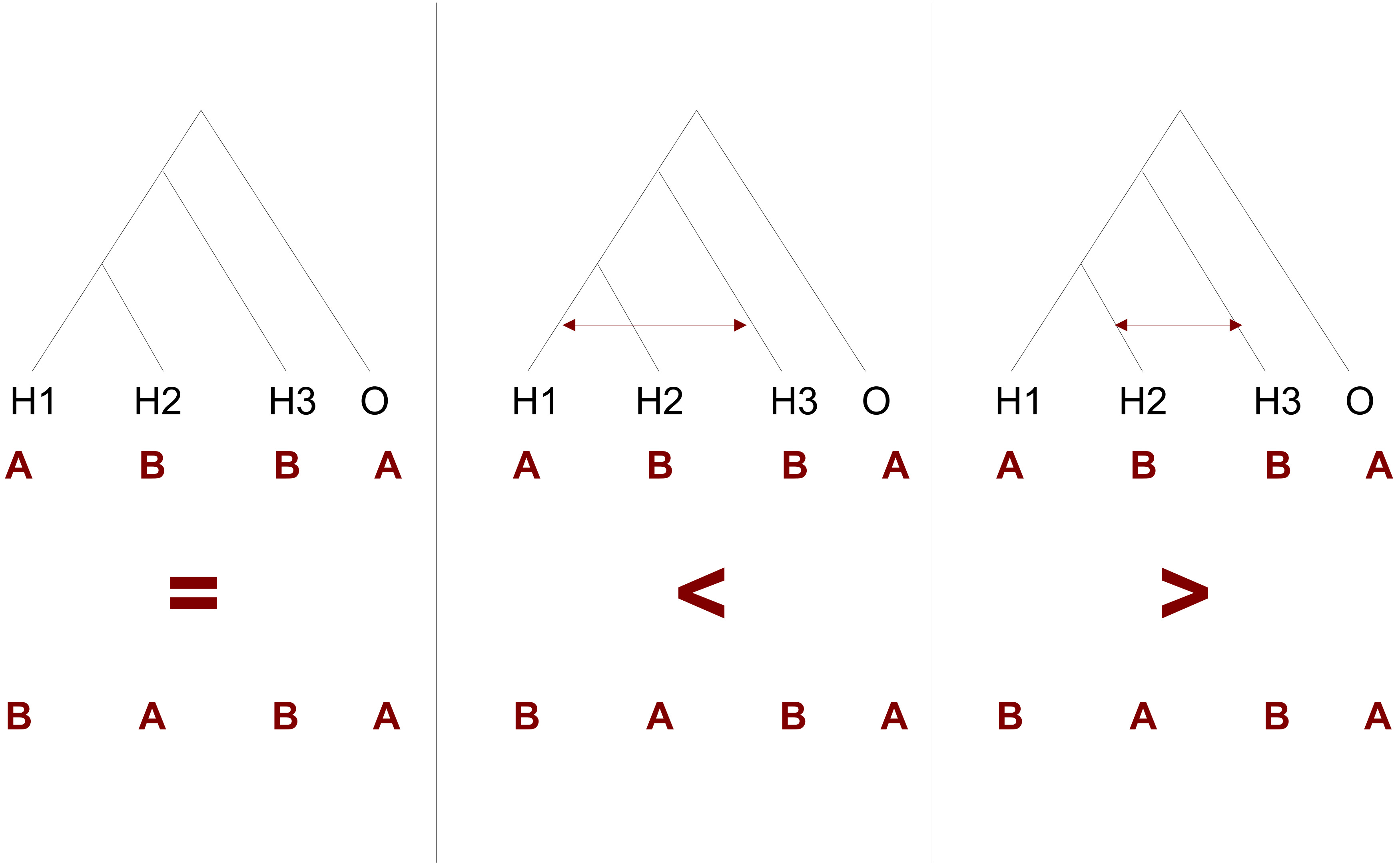
F4 statistic and testing for treeness



What kind of violations of this tree would affect the F4 statistic?

Are there any kind of admixture/migrations that would make the F4 statistic weird?

D statistic (ABBA-BABA)



H1 H2 H3 O

A B B A

=

B A B A

H1 H2 H3 O

A B B A

<

B A B A

H1 H2 H3 O

A B B A

>

B A B A

Exercise time!!!